Title

Models, Measurement, and Metrology Extending the SI: Trust and Quality Assured Knowledge Infrastructures

Permalink

https://escholarship.org/uc/item/2j73n8mk

ISBN

978-3-11-103649-6

Publication Date

2024-09-23

DOI

10.1515/9783111036496

William P. Fisher, Leslie Pendrill (Eds.)

Models, Measurement, and Metrology Extending the SI

De Gruyter Series in Measurement Sciences

Edited by
Klaus-Dieter Sommer and Thomas Fröhlich

Models, Measurement, and Metrology Extending the SI

Trust and Quality Assured Knowledge Infrastructures

Edited by
William P. Fisher and Leslie Pendrill

DE GRUYTER
OLDENBOURG

Editors
Dr. William P. Fisher
BEAR Center
Berkeley School of Education
University of California, Berkeley
1936 University Ave. Ste 355
Berkeley, CA 94704
wfisher@berkeley.edu

Dr. Leslie Pendrill
Storskiftesv. 11B
SE-433 41 Partille
Sweden
leslie.pendrill@ri.se

ISBN 978-3-11-103623-6
e-ISBN (PDF) 978-3-11-103649-6
e-ISBN (EPUB) 978-3-11-103715-8
ISSN 2510-2974
DOI https://doi.org/10.1515/9783111036496

Library of Congress Control Number: 2024937642

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the internet at http://dnb.dnb.de.

Cover image: *Wassily and the World of Metrology*, Susanna Beyer, llmenau, Deutschland Aus der Sammlung: Measurement lmpressions {Privatbesitz)
Typesetting: Integra Software Services Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Psychometric Foreword

There has been substantial work carried out in the last decade investigating the intersection of measurement in the physical sciences with measurement in the human sciences. For example, Mari et al. (2023) have extended a claim in the title of their book: *Measurement Across the Sciences: Developing a Shared Concept System for Measurement*. Yet, they also note,

> Invoking the language of measurement connotes *epistemic authority*: measurement has historically been associated with epistemic virtues such as objectivity, precision, accuracy, and overall trustworthiness, largely as a result of the highly successful history of measurement in the physical sciences and engineering. But, *prima facie*, it is not clear whether measurement processes outside these fields actually deserve to be associated with such authority. (p. xx)

The broad notion of this collection of chapters is that one way to engage with this question is by examining the possibilities (i.e., "viability, feasibility, and desirability," from the title of Chapter 1) for extending the SI into the human sciences space. The chapters herein range across a wide scope of measurement properties ("attributes"), measurement objects, and measurement issues. As such they constitute multiple instances of a sort of "proof of life" for this supposition.

Note that the chapters do not necessarily establish the viability of that "life" beyond the confines of their own application areas. But, indeed, one can see the possibilities sparking out from among them. This volume will be seen as a provocation by some, and an invitation by others, to take on the foreseeing of an extended SI, one that will likely require both physical scientists and human scientists to "drive this wedge between scientific objectivity and hermeneutic relativity" and hence avoid the assumption that "whatever needs to be interpreted in order to be understood will, to that extent, become a matter of taste or subjectivity" (Toulmin, 1982, p. 94).

I invite the reader to undertake this journey with the chapter authors and expand the range of appreciation for measurement, its beauty, and its immense scope.

Mark Wilson
Distinguished Professor
Director, BEAR Center
Berkeley School of Education
University of California, Berkeley

References

Mari, L, Wilson, M. & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement, Second edition.* New York: Springer.

Toulmin, S. (1982). The construal of reality: Criticism in modern and postmodern science. *Critical Inquiry, 9* (1), 93-111.

https://doi.org/10.1515/9783111036496-202

Metrological Foreword

The *De Gruyter Series in Measurement Science* (*DGSM*) includes monographs ranging from the mathematical foundations of metrology, the link between metrology and information theory, and dynamic measurements to recent developments such as quantum sensing and cognitive sensors and measurement systems.

The present volume "Models, Measurement and Metrology Extending the SI," has been produced under the leadership of the internationally renowned metrologists *Wiliam P. Fisher* and *Leslie Pendrill.* It is the third of the monographs in the DGSM series. The other planned volumes in the series will be available within a period of 24 months after the publication of this volume.

Today, trust and quality assured knowledge are given by the International System of Units (SI). It enables all physical quantities to be referred to with one or more of the seven defining constants. Almost all quantities are traced back to high precision quantum standards. This means that all measured values are unambiguously comparable when they refer to calibrated artifacts. The quality and trust of a measured value are usually expressed by the measurement uncertainty. The rules and terminology for traceable measurements are internationally defined and globally accepted. This self-contained system enables the comparison of measured values and therefore forms a central backbone for trade, science, and society.

In this book titled “Models, Measurements and Metrology extending the SI,” the authors present a broad spectrum of measurement properties, measurement objects and measurement problems in psychology, vision science, and the social sciences. They discuss questions and name initial prerequisites that are necessary to extend the SI for these applications. Here, definitions of traceable references are much more challenging to identify and estimate as compared to physical quantities, as we need to rethink concepts such as objectivity, precision, accuracy and trust when they are based on observations of human behaviour, on personal opinions and understandings, or on higher order aggregations of physical measurements, as in *Massof's* chapter on modeling an equal entropy scale for measuring the dry eye disease state.

Large variations exist in the quality of what are commonly referred to as measurements in psychology and the social sciences. Widely accepted methods referred to as “quantitative” illogically treat ordinal scores as though they are interval quantities. Statistical comparisons are made even when no unit is defined, and the meaning of numeric differences not only changes with scale locations but, moreover, depends entirely on the characteristics of the local respondent or examinee sample and the particular instrument used. These issues raise fundamental questions about the potential for a reference system of robust, globally accepted standards, and, above all, how comparable measurements can be carried out and validated. These fundamentals are the focus of the chapters in this volume.

https://doi.org/10.1515/9783111036496-203

We invite readers to undertake this journey into the future of metrology together with the chapter authors, and to thereby expand their range of appreciation for measurement, its beauty, and its immense scope.

Klaus-Dieter Sommer
Technische Universitaet Ilmenau (Germany)
Editor of the DGSM Book Series

Frank Haertig
Vice President of the Physikalisch-Technische
Bundesanstalt (Germany)
President of the International Measurement
Federation (IMEKO)

Contents

Part I: **Theory and Principles in Measurement and Metrology**

Part II: **Designing and Calibrating Metrologically Viable Measurements: Methods and Applications**

William P. Fisher Jr. and Leslie R. Pendrill

1 Introduction: imagining the viability, feasibility, and desirability of extending the SI to include the psychological and social domains

Abstract: Metrology – quality-assured measurement – provides a commonsense way of connecting scientific and mathematical thinking with everyday thinking. Today's predominant approaches to quantification in psychology and the social sciences are inadequate to the urgent challenges humanity faces. Instead of hollow imitations of the measurement methods employed in physics, a metrological perspective grounds measurement across the sciences in cognitive and social processes familiar to all. Instead of foregrounding quantification as the primary determinant of measurement thinking, a more productive path forward orients social and psychological measurement toward the same sources in everyday thinking that were extended into science by physicists and engineers. In this introduction to *Models, Measurement, and Metrology Extending the SI: Trust and Quality Assured Knowledge Infrastructures*, we take up misconceptions about measurement, describe the chapters in this volume, summarize the history of developments in measurement theory, recount some recent social history, and generally pursue some implications of the metrological shift in perspective.

Keywords: psychological measurement, measurement modeling, metrological traceability, history of measurement, philosophy of science, psychometrics, Rasch models, quality assurance

1.1 To begin

As an introduction, our aim is to describe a metrological frame of reference for the other chapters in this book. A metrological context for psychological and social measurements will certainly sound, at the very least, counterintuitive to many readers. The very notion blatantly contradicts the analytical, statistical, and ordinal methods of quantification associated with achievement tests, surveys, rating scales, and assessments of various kinds.

Acknowledgment: We would like to thank the University of California, Berkeley Libraries and multiple chapter authors' independent funding sources for their support in making the book open access.

William P. Fisher Jr., BEAR Center, Berkeley School of Education, University of California, Berkeley; Living Capital Metrics LLC

Leslie R. Pendrill, Metrology, RISE Research Institutes of Sweden

https://doi.org/10.1515/9783111036496-001

Metrology – which has been formulated predominantly in the physical sciences – structures measurements read from instruments calibrated to quality-assured interval or ratio unit standards and disseminates traceability throughout networks of end users in support of interoperability. In psychology and the social sciences, in contrast, quantitative methods have traditionally focused on centrally planned and controlled ordinal data gathering and analysis. That shift in context of metrology to include psychological and social domains requires some elaboration, as many readers will not only be unfamiliar with it, they understandably may also find it inappropriate, misleading, or misguided.

This chapter briefly addresses some common misconceptions and unexamined preconceptions about measurement and metrology. This not only will draw out explications of longstanding and proven technical capabilities but will also highlight associated social and moral capabilities that must be included in any account of this kind. These latter are not rationalizations added on after the fact as ad hoc justifications for a predetermined purpose but are integral to the logic, aesthetics, and ethics of measurement. Measurement is, after all, deeply rooted in social contexts requiring a shared sense of fair dealing, such as trade and taxation. Although these norms usually remain in the background as unexamined assumptions, they nonetheless influence and shape the form of the agreements and distinctions we make (Alder, 2002, p. 2). In this respect, the standards we set and accept as guidelines for respectable or criminal behavior define what we value. This book takes up the problem of whether the standards informing our governance, market, educational, health care, and environmental institutions today adequately represent our cultural self-image, or if perhaps we could do better at living up to the ideals represented by the symbolic scales of justice.

These considerations point us toward a key insight, namely, that “the development from the spoken language . . . to symbols and pictograms . . . to what we now understand as written language is a perfect standardization process” (Weitzel, 2004, p. 11). Taking the development of written language as a model for an extended SI affords some key advantages. After documenting a detailed history of various efforts aimed at improving the human condition, Scott (1998, p. 357) recommends language as providing the best model for adaptable human institutions. He makes this recommendation because language is “a structure of meaning and continuity that is never still and ever open to the improvisations of its speakers.” The fact that the existing SI undergoes continuous improvements while remaining in uninterrupted use speaks to its continued performance as an extension of everyday language.

To succeed, an extended SI must also possess exactly that kind of globally navigable comparability at the same time it accommodates local adaptations. The need for evolutionary potential in an extended SI will arguably be even more important than in the existing SI. This potential is realized by the demonstrated capacities of probabilistic models of measurement. Those models support empirical and theoretical estimations of measured quantities exhibiting persistent and reproducible invariances (within the range of uncertainty) across the shifting memberships of human populations and the contents of tests, item banks, and equated instruments. Measurement modeling in this context must encompass the complex demands of evolving circumstances, as the people involved and the challenges they face constantly change. It must also provide a context in which unique individuals can experience the powerful affirmation of feeling part of something larger than themselves without at the same time being overwhelmed by it. Given developments in psychological measurement theory and practice over the last 60 years and more, we feel confident this can be accomplished and has indeed been accomplished repeatedly over the last several decades.

This introduction cannot, of course, address the topic raised in anything approaching a comprehensive manner. Indeed, an entire book or series of books would be insufficient for that task. More fundamentally, one of the essential themes involved in this domain concerns the fact that the value of metrologically informed measurement is not, and cannot be, communicated primarily via merely cognitive forms of understanding. Lived experiences involving personal engagement, emotional associations, embodied sensations, economic and political consequences, creative opportunities for innovation, etc. will be required before the ideas offered here can be properly understood and evaluated.

Finally, though many positive benefits may be expected to accrue from new metrological infrastructures affording the mass distribution of higher quality and comparable forms of information, we do not regard the challenges involved as simple or the solutions as panaceas. We fully expect extending the SI into psychology and the social sciences to be controversial and difficult, and resource intensive. In this, we accept the lessons of history, recognizing that many of the original goals set over 200 years ago for the West's original unified system of decimalized metrics had to be abandoned or significantly modified. Those lessons also show that significant returns on investment were provided on massive up-scaling. The successes of those efforts have, however, altered the environment in which they occurred to such an extent that instead of enhanced prosperity they now threaten catastrophe. New ideas more explicitly cognizant of the organism-environment unity as the focus of natural selection are in order.

1.2 An audacious but grounded proposal: extending the SI

Given the usual understanding of quantitative methods as assigning numbers to operations, or more colloquially, reducing things to numbers, many readers will likely find the idea that the SI could be usefully, meaningfully, ethically, and productively extended to apply to human affairs, a preposterous, and even a dangerous, proposition. Given the typical perception that the natural world is measurable because it is inherently and automatically numerical, to many the notion of an extended SI applied to human subjectivity will sound positively ridiculous and not worth considering.

Readers with some background in the history and philosophy of science will have a somewhat – in some cases, a markedly – different perspective. This will especially be the case for those involved in science and technology studies, conceptual metaphor theory, cognitive and developmental psychology, and related areas. Even for these readers, however, the complexities of the issues may seem intimidating. Some few may find the idea that the SI could plausibly be extended into human affairs to invite curiosity.

Those readers familiar with technical developments in measurement theory and practice, finally, may be breathing a sigh of relief that the topic has finally been broached in a way that positions it as a matter for broad and serious consideration, and not as something to be dismissed out of hand. After all, the mathematical identity of measurement models applicable across the natural and social sciences was deemed "widely accepted" since the 1960s by two noted authorities almost 40 years ago (Narens & Luce, 1986).

One particularly lethal misconception about measurement must be immediately addressed: measure equations and quantity equations are not to be confounded; they are distinct concepts (Maxwell, 1873; de Courtenay, 2015). As can easily be shown, measurement is at the same time not fundamentally or primarily a matter of quantification (Mari et al., 2013, 2016). Many scientists, philosophers, and psychologists note that the inferential processes involved in mathematical thinking and quantitative modeling are already present in everyday language and thinking (Black, 1946, pp. 304–305; Bohr, 1963, p. 9; Einstein, 1954, p. 290; R. Fisher, 1935, p. 79; Guttman, 1994; Huxley, 1862, pp. 57–58; Nersessian, 2008). Metaphysically speaking, the intuition that the universe is somehow numeric in and of itself increasingly seems to be an incomplete and undeveloped initial insight into the idea that existence is broadly mathematical, and is so in profoundly logical, beautiful, and ethical ways. Understanding this does not require a high-level technical understanding or a convoluted esoteric philosophy. A commonsensical point of view is offered throughout this book, one that leads to end results devoid of hairsplitting logical arguments of limited applicability. Our arguments and those of the other authors of the chapters in this book instead open an array of highly pragmatic opportunities for the advancement of the arts and sciences and associated prospects for cultural and economic progress.

A hint that the SI could profitably be extended to include psychological and social domains might be found in the repeated observations of multiple measurement wheels reinvented *ad nauseam* in those areas over the last several decades. The reproduction of linearly comparable quantities across different studies employing different instruments on different samples was, in general, highly unexpected and not predicted by the researchers involved. But persistently surprising convergences keep coming to light. How are these results being produced? How can they be explained? Could they possibly be systematically developed into a general frame of reference for scientific methods? What might they mean for improved understanding and communications? What do they imply for social organization, and the relation of human beings to the Earth and the world of nature? To begin answering these questions, we should take a moment to retrace our steps and consider how we arrived at this point. We will focus in turn on issues of meaningfulness, economic incentives and rewards, and first philosophy, or metaphysics, after first giving an overview of the chapters in the book.

1.3 The chapters

The chapters in this book demonstrate how interval and ratio level measurements of social and psychological constructs are in principle capable of comprising an extended SI. This capability follows from the definition and estimation of unit quantities that retain their properties independent of particular samples and instruments. Quantitative measurements of this kind have been produced using well-established probabilistic models and methods for well over 50 years. The following chapters illustrate how to employ advanced measurement modeling to obtain one or more of several advantageous affordances.

Part I of this volume presents various chapters under the heading of Theory and Principles in Measurement and Metrology, while Part II takes up Designing and Calibrating Metrologically Viable Measurements: Methods and Applications. Leslie Pendrill (2024) sets the stage by following through from historical developments in psychology to present-day foundational studies of the applicability of traditional engineering of measurement systems and concept systems (quantities, units and relations between them) to human-based metrology. Matthew Barney and Feynman Barney (2024) then envision an integration of metrology, psychological measurement models, and artificial intelligence, describing how common metrics might be embedded in everyday life via quantities unobtrusively inferred from decisions and behaviors. Ernesto San Martin and colleagues (2024) explore the value of identified and partially identified models for inferring causal relationships in measurement. Jeanette Melin (2024) expands on the theme of causality in her investigation of validity as a matter of essential concern in an extended conceptualization of the SI. William P. Fisher, Jr. (2024) concludes Part I with reflections on some of the paradoxical reasons why metrological traceability to unit standards is not typically considered possible in social and psychological measurement, and how it is also eminently reasonable and urgently needed.

Part II's examples of metrologically viable measurement scaling begin with a contribution from Robert Massof, Chris Bradley, and Allison McCarthy (2024) showing how clinical signs and symptoms can be organized to model a continuous disease state variable. Steve Lang and Judy Wilkerson (2024) follow this with a description of their journey in the development of an affective assessment for measuring teacher dispositions, a domain not nearly as well investigated as that of cognitive or behavioral constructs. Harry Kollias (2024) contends with the often life-changing consequences of high-stakes assessments involving panels of judges who vary in their perceptions of task mastery and the satisfaction of performance standards. Greg Sampson and colleagues (2024) describe another high-stakes assessment context involving similar multifaceted problems in the licensure and certification of emergency medical technicians. Simon Karlsson and colleagues (2024) address municipal sustainability metrics with the overt intention of capitalizing on the metrological potentials of a distributed system of manageable metrics.

Trudy Mallinson (2024) asserts that "measurement is not a benign act" but has consequences for equity and social justice that demand close attention in devising and applying justice-oriented anti-racist criteria for validating quantitative results. Linda

Morell, Sean Tan, and Mark Wilson (2024) document the interrelationships of assessment content and inferences as to competencies in the measurement of twenty-first century skills. Dhanya Natha Kumar and Hrishi Joshi (2024) focus on surgical outcomes assessments and how they could better contribute to the advancement of patient-centered care if they were designed, calibrated, and maintained with closer attention to fundamental measurement principles. Finally, David Sul (2024) takes up the question as to how measurements need not elevate any single cultural worldview to a position from which it is allowed to negate or erase the values of any other worldview.

In different ways, all these chapters show how a meaningful extension of the SI to cover the psychological and social domains depends on and benefits from the following conceptual distinctions and methodological demands. Readers unfamiliar with technical issues in measurement and metrology should approach the chapters with the following in mind and might be motivated to explore these ideas at greater length if practical applications are to be undertaken:

- Ordinal scores are not interval measurements, just as numeric counts are not measured quantities (Wright, 1992b; Wright & Linacre, 1989). Everyone knows we cannot say who has more rock when I have two and you have five, yet we persist in fallaciously treating test scores as measurements in the absence of a defined unit quantity. Even though psychology and the social sciences have for several decades possessed the models and methods needed for defining and estimating quantities from counts, global institutions continue to assume counts and percentages of correct answers and of ratings suffice as quantitative measurements. This book argues that there are firm reasons based in theory and evidence for altering that circumstance.
- Stevens' (1946) initial categorization of four levels of measurement (nominal, ordinal, interval, and ratio) was later augmented by a fifth log-interval level (Stevens, 1957, 1959) recognized by Narens and Luce (1986) as in productive use across the natural and social sciences and in applications of additive conjoint measurement models like those described in this volume. In a recent review, Marmor and Bashkansky (2020) consider adding new types of quality data to the well-known nominal, ordinal, interval, and ratio scales, including for example ranked data, as previously proposed by Tukey (1986).
- Ordinal score data volume is reduced by interval measurement with no loss of information, as individual observations are both necessary and sufficient to the estimation of measurement model parameters (Andersen, 1977, 1999; Fischer, 1981).
- Physical measurements of length and distance, mass, and density expressed in SI units have been reproduced from ordinal observations via the application of probabilistic conjoint measurement models (Choi, 1998; Moulton, 1993; Pelton & Bunderson, 2003; Stephanou & Fisher, 2013).
- The reproducibility of patterns of concrete observations from abstract measurements and formal theory is never perfect, of course, so data displays revealing failures of invariance pertinent to end user interests in improved outcomes are of value in

bringing out qualitative exceptions to the rule (Allen & Pak, 2023; Chien et al., 2011, 2018; Fisher, Oon, & Benson, 2021; Wright & Stone, 1979; Wright et al., 1980).

- Substantive integrations of qualitative and quantitative data and methods in formative assessments delineate learning progressions and map variation to locate individuals or groups relative to where they have been, where they want to go, and what to do next (Black et al., 2011; Duckor & Holmberg, 2019; Morell et al., 2017).
- Repeatable and reproducible definitions of meaningful additivity are essential to models sufficient to the identification of constructs as independent, within the range of uncertainty, of the sample measured and the instrument used to measure (San Martin & Rollin, 2013; San Martin et al., 2024; Wright, 1997a/b).
- Despite persistent repetitions of the idea that Rasch's models for measurement are one-parameter IRT models, they have no connection whatsoever with item response theory (Wright, 1984, 1997b). As Cliff and Keats (2003, p. 15) recognized, it was only "by means of some highly dubious assumptions [that] the Rasch formulation was generalized to what is now called Item Response Theory."
  - Neither Rasch nor any of the major contributors to the development of measurement theory and practice based in Rasch's models assert a connection with IRT.
  - Despite the inclusion of a model equation of the same form as those described by Rasch, the meaning of the parameters is entirely different.
  - IRT rationalizes unidentified and internally contradictory item parameters in an overt prioritization of a modern, positivist focus on describing data presumed to exist independent of human interests.
  - This renders IRT constitutionally incompatible with measurement theory, as has long been asserted (Andrich, 1989; Cliff, 1992, p. 188; Embretson, 1996a; Lumsden, 1978, p. 22; San Martin et al., 2015; Verhelst & Glas, 1995, p. 235; Wood, 1978, p. 31).
- That the association of Rasch's models with IRT persists despite repeated explanations of the illogic involved may be another example of "effortless associative thinking" (Kahneman, 2003, 2011; Simon, 1997, 2000) characteristic of bounded rationality and exemplifying the powerful constraints imposed on thinking by the dominant paradigm:
  - Weiss (2021), for example, despite having "led the much of the seminal research behind Computerized Adaptive Testing, and trained several generations of eminent psychometricians," nonetheless
    - continues to categorize Rasch's models in IRT,
    - holds that Rasch's models have been "been replaced by more general [IRT] models that allow test items to vary in discrimination, guessing, and a fourth parameter,"
    - considers the models' invariance requirement to be an expendable assumption, and

-
  -
    - concludes that Rasch's models "can best be viewed as an early historical footnote in the history of modern psychometrics."
  - This modernist IRT perspective is then inherently at odds with an unmodern metrological perspective, both in theory and in the facts of the continued rapid growth and development of Rasch-based theory and practice in the field (Aryadoust et al., 2019).
  - We may then have here an instance in which statistically oriented and measurement-oriented communities of research and practice will have to agree to disagree:
    - In the spirit of the "unified disunity" and "convergent divergence" (Blok et al., 2016, 2020; Bowker et al., 2015; Galison, 1997, 1999; Woolley & Fuchs, 2011) in thinking ineradicable from irreducible complexity, there is more to be gained from passionately held differences of opinion than from hollow and coerced consensus.
    - Of course, the unmodern paradigm does not eliminate metaphysics or bounded rationality from playing roles in science; it can only expand the limits in which "effortless associative thinking" obtains.
    - When it does, and new ecological economies of thought transform today's scientific, legal, market, and communications infrastructures, few will likely quibble over differences that do not make a difference.
- Multiple specialized software packages are available for testing empirical and theoretical invariance hypotheses and for estimating measurement and calibration parameters, for a wide variety of models (Adams et al., 2020; Andrich et al., 2017; Bulut, 2021; Doran et al., 2007; Hohensinn, 2018; Lamprianou, 2020; Li, 2006; Linacre, 2023b, 2024; Melin & Pendrill 2023; Pendrill 2024; Robitzsch et al., 2020; Torres Irribarra & Freund, 2014; Verhelst et al., 2007; Wilson et al., 2019).
- Comparisons of estimation algorithms and program outputs are available (Linacre, 2023a; Linacre et al., 2013; Robinson et al., 2019; Yumoto & Stone, 2011).
- The Rasch.org website offers the full texts of authoritative books, *Rasch Measurement Transactions*, and articles; information on software training workshops, and conference calendars.
- Rigorous separation of levels of complexity, where formal construct theory explains variation in abstract item calibrations and person measurements estimated from concrete observations, enable theoretical, metrological, and experimental issues to be dealt with by separate, collaborative communities of research and practice.
- Explanatory models that successfully predict and account for large proportions of variation in instrument calibrations and sample measurements make it possible to automate item generation on the fly and to infer ratings and measurements from observed behaviors (Barney & Barney, 2024; De Boeck & Wilson, 2004; Embretson, 2010; Fischer, 1973; Stenner et al., 2013) and in some cases can even enable metrological references for traceability (Pendrill, 2019, 2024; Melin et al., 2021).

– Finally, and perhaps most counterintuitively to many readers, culturally specific assessments can measure abilities and attitudes in terms meaningful to varied communities without compromising broader capacities for comparability (MacIntosh, 1998; Mallinson, 2024; Sul, 2024; Tennant et al., 2013; Teresi et al., 1995; Wilson, 1994a).

1.4 Some background history

1.4.1 Origins

L. L. Thurstone was a former electrical engineer who went into psychology, was the first president of the Psychometric Society, was a co-founder of *Psychometrika*, and was the director of examinations at the University of Chicago from 1924 to 1952. In 1928, he wrote:

> The scale must transcend the group measured.–One crucial experimental test must be applied to our method of measuring attitudes before it can be accepted as valid. A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is so affected, the validity of the instrument is impaired or limited. If a yardstick measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement. (Thurstone, 1959, p. 228)

Introducing the conception and purpose of the measurement models he devised, on the first page of his 1960 book, Rasch expanded on this theme, saying:

> Individual-centered statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated. It is further essential that comparisons between individuals become independent of which particular instruments – tests or items or other stimuli – within the class considered have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class – 'measuring the same thing' – independent of which particular individuals within a class considered were instrumental for the comparison. (Rasch, 1960, p. xx; also see Rasch, 1961, 1966a/b)

Rasch then intentionally formulated an individual-level model for measurement structured on the basis of observations of any kind of physical or psychological phenomenon:

> taken as nothing more than an accidental response, as it were, of an object – a person, a solid body, etc. – to a stimulus – a test, an item, a push, etc. – taking place in accordance with a potential distribution of responses – the qualification 'potential' referring to experimental situations which cannot possibly be [exactly] reproduced. (Rasch, 1960, p. 115)

Rasch showed that such response distributions “depended on one relevant parameter only,” one chosen so that the same multiplicative law applied no matter whether observations involved people, solid bodies, test items, or pushes. He concluded that:

> Where this law can be applied it provides a principle of measurement on a ratio scale of both stimulus parameters and object parameters, the conceptual status of which is comparable to that of measuring mass and force. Thus, . . . the reading accuracy of a child . . . can be measured with the same kind of objectivity as we may tell its weight. (Rasch, 1960, p. 115)

Four years later, in 1964, Luce and Tukey showed how:

> the fundamental character of measurement axiomatized in terms of concatenation [is extended to] qualitatively described ‘additivity’ over pairs of factors of responses or effects . . . [such that] the additivity is axiomatizable in terms of axioms that lead to scales of the highest repute: interval and ratio scales. (Luce & Tukey, 1964, p. 4)

They illustrated that measurement axiomatized on the basis of conjointly ordered pairs of factors “apply naturally to problems of classical physics and permit the measurement of conventional physical quantities on ratio scales,” concluding that:

> In the various fields, including the behavioral and biological sciences, where factors producing orderable effects and responses deserve both more useful and more fundamental measurement, the moral seems clear: when no natural concatenation operation exists, one should try to discover a way to measure factors and responses such that the ‘effects’ of different factors are additive. (Luce & Tukey, 1964, p. 4)

In 1986, Narens and Luce, then, generalized from examples spanning different physical and behavioral contexts (“the ordering by mass of objects characterized by their volume and density; the loudness ordering provided by a person for pairs of sounds, one to each ear; and the preference ordering provided by an animal for amounts of food at certain delays”). They showed that conjointly ordered effects of these kinds:

> not only provided a deep measurement analysis of the numerous nonextensive, ‘derived’ structures of physics, but also provided a measurement approach that appears to have applications in the nonphysical sciences and has laid to rest the claim that the only possible basis for measurement is extensive structures. (Narens & Luce, 1986, p. 177)

On the basis of these results, Narens and Luce (1986, p. 169) say that, with the introduction of the theory of additive conjoint measurement in the 1960s (Andersen, 1970; Brogden, 1977; Fisher & Wright, 1994; Green, 1986; Luce, 1959, 1978; Luce & Tukey, 1964; Newby et al., 2009; Pelton & Bunderson, 2003; Perline et al., 1977; Rasch, 1960, 1961, 1966a/b; Wright, 1968, 1977), the view that interval-scalable, fundamental measurement is possible for nonextensive structures became “widely accepted.” Andrich (1988, p. 22) concurred, pointing out that, “. . . when the key features of a statistical model relevant to the analysis of social science data are the same as those of the laws of physics, then those features are difficult to ignore.”

Despite the supposed wide acceptance of interval scalable results, these models remain far from defining mainstream measurement theory and practice. Wright (1968, 1977, 1997a/b, 1999; Wright & Masters, 1982; Wright & Stone, 1979, 1999) brought Rasch's ideas into fairly wide use, emphasizing that his models for measurement are not data models but are definitions of the laws of measurement (Wright, 1988). Wright and his colleagues and students (Bode et al., 2000; Chien et al., 2011; Connolly et al., 1971; Kielhofner et al., 2005; Linacre, 1997; Liu, 2018; Massof, 2008; Mead, 2009; Smith, 1994, 1997; Stenner et al., 2013; Wolfe et al., 2000; Wright, 1997a, 2012; Wright & Stenner, 1998; Wright & Stone, 1979) repeatedly addressed problems related to the definition of meaningful units and the design of instruments, applications, and reports incorporating them.

Wright (1997a, 2012; Wright et al., 1980; Wright & Stone, 1979) thought to leverage the inferential stability of established calibrations by making measurement and uncertainty estimates, along with graphical response consistency evaluations, available at the point of use as soon as observations were recorded. Given constructs proven via explanatory theory and experimental evidence to be stable across multiple samples and instruments, the next case through the door would not likely alter the definition of what was measured. Instead of assuming measurement is only ever a product of data analyses, repeated empirical validations of theoretical predictions afford the opportunity to devise self-scoring forms interpretable at the point of use. Wright may never have once included the word "metrology" in his writing, but he nonetheless articulated a fundamentally metrological goal when he saw how on-the-spot applications could be supported by measurements read from quality-assured instruments calibrated in a reference standard quantity.

In so doing, with no reference to the history of science, what Wright understood about measurement as the actionable modeling of the real world was well described by Ackermann (1985, pp. 143–144) when he noted that:

> Once clear statements of fact have been achieved through instrumental investigation, the reference of fact seems fixed and objective, and indeed it is. The world has been discovered to show a fixed and repeatable response in certain interactions as described in the language, and this response is an objective consequence of these interactions. . . . This process of achieving or constructing reference for language by development of a domain we will call the microprocessing of fact, after discussion of this phenomenon by Latour and Woolgar [1979]. When the process is complete, the evidence of microprocessing disappears, and mere correspondence, the very correspondence that has been slowly and carefully constructed, is all that remains.

Wright's admonitions and recommendations in this regard are not often followed, but the value of the probabilistic models of measurement he advocated was recognized by many soon after they were introduced (Wilson & Fisher, 2017). These models have been further explicated and increasingly applied in psychology, health care, and the social sciences over the last several decades (Andersen, 1977, 1980; Andrich, 1978, 2010; Andrich & Marais, 2019; Aryadoust et al.; 2019; Bezruczko, 2005; Boone & Staver, 2020; Embretson, 1996b, 2010; Engelhard, 2012; Fischer, 1973, 1981; Fischer & Molenaar, 1995; Fisher &

Wright, 1994; Hagell, 2014, 2019; Loevinger, 1965; Massof, 2008; Masters & Keeves, 1999; Pendrill, 2019, 2024; Melin et al., 2021; Pesudovs, 2006, 2010; Salzberger, 2009; Smith & Smith, 2004, 2007; von Davier & Carstensen, 2007; Wilson, 1992, 2018, 2023).

After expanding on developments in measurement theory dating back several centuries, including the introduction of Rasch's, and Luce and Tukey's, additive conjoint perspectives, Wright (1997b, p. 44) concludes his history of social science measurement saying that "Today there is no methodological reason why social science cannot become as stable, as reproducible, and hence as useful as physics." Although Wright (1997b, p. 33) recognized the social roots of uniform metrics in society's demands for fair taxation and trade, and though he clearly stated that "Science is impossible without an evolving network of stable measures," he did not reflect on the relevant metrological challenges or opportunities. O. D. Duncan, in contrast, "the most important quantitative sociologist in the world in the latter half of the twentieth century" (Goodman, 2007, p. 131), articulated a metrological connection with Rasch's models for measurement. Duncan then plays a key role in expanding on Wright's claim as to the methodological potential for social science to become as stable, reproducible, and useful as physics.

1.4.2 Otis Dudley Duncan's contributions

Over the course of the 1980s, Duncan introduced Rasch's measurement models into sociology. In so doing, Duncan (1984b, pp. 38–39) suggested "that social measurement should be brought within the scope of historical metrology, while that discipline learns to take advantage of sociological perspectives." Toward that end, Duncan argued that:

> What we need are not so much a repertoire of more flexible models for describing extant tests and scales . . . but scales built to have the measurement properties we must demand if we take 'measurement' seriously. As I see it, a measurement model worthy of the name must make explicit some conceptualization – at least a rudimentary one – of what goes on when an examinee solves test problems or a respondent answers opinion questions; and it must incorporate a rigorous argument about what it means to measure an ability or attitude with a collection of discrete and somewhat heterogenous items.
>
> Thurstone explicated the meaning of measurement as it might be accomplished by such an instrument. Rasch provided the formalization of that meaning. (Duncan, 1984b, p. 217)

Complementing that view on models implementing a rigorous conception of measurement, Duncan (1984b, pp. 206–207) gives a long list of "ambiguous and poorly discriminated concepts" as evidence of "the prevailing chaos in which there is a multiplicity of 'tests,' 'scales,' or 'instruments' ostensibly serving as 'measures'" but which fail to live up to even a generous sense of the word. He cites research showing:

> many instances of the same items (questions, or statements calling for an agree/disagree response) in tests intended to measure different constructs, different and dissimilar items in tests

> with the same or similar names, a widespread habit of arbitrarily modifying tests when applying them in new research (and thereby precluding comparison or any benefit of standardization), and the replacement of old scales by new ones without cross-calibration between them and without demonstration of improved validity. (Duncan, 1984b, p. 207)

The phenomenon of myriad incommensurable metrics was also once the case in physics, as Duncan brings out in his notes on historical metrology and has also long been amply demonstrated in the history of science (Alder, 2002; Ashworth, 2004; Black, 1962; Crosby, 1997; Hesse, 1970; Kula, 1986; Nersessian, 2008; Roche, 1998; Wise, 1995).

In this context, somewhat inadvertently, Duncan is here offering evidence that the dominant modern paradigm's conception of science as describing an independently given objective reality is fatally flawed. The question is, what methodological conclusions might we draw from the historical coevolution of metric standards with concurrent developments in conceptually aligned governance and economic principles, with the co-production of science and society (Bowker, 2016; Edelmann, 2022; Ihde, 1991; Jasanoff, 2004; Knorr Cetina, 1999; Power, 2004)?

The reductionist conception of quantity sees it as built up from elementary building blocks in the physical world. Here, wholes are the sums of parts. The associated idea of measurement as only describing a pre-existing given reality not only fails to hold in the history of the natural sciences but also undermines the very methodological foundations of psychology and the social sciences. To be sure, there is a marked and highly relevant distinction to be made between this naïve sense of an inherently quantitative universe and the objectively repeating and reproducible self-organized phenomena that persistently exhibit consistent properties across samples, instruments, laboratories, observers, time, and space. The point to be made is limited to noticing that uncritically held metaphysical faith in a transcendent reality falsely makes it appear that the incommensurability of metrics in sociology and psychology is a consequence of human subjectivity disconnected from objective reality. It is nothing of the sort.

Duncan's description of the chaos in social measurement indirectly amplifies the point made by Gödel (1931) and a wide range of others (Garfinkel, 1991; Lerner & Overton, 2017; Nagel & Newman, 1958; Toulmin, 1953; Wittgenstein, 1983) as to the existence of arithmetical truths that cannot be formally demonstrated. Gödel thought, and many others have agreed, that his proofs of this theorem ought to have been elevated to a fundamental principle of science equivalent to Einstein's theory of relativity (Calude, 2002, 2007; Chaitin, 1994; Floyd & Kanamori, 2016). In this vein, Holton points out that

> we can find even among the most hard-headed modern philosophers and scientists a tendency to admit the necessity and existence of a noncontingent dimension in scientific work. Thus Bertrand Russell speaks of cases 'where the premises of sciences turn out to be a set of presuppositions neither empirical nor logically necessary'; and in a remarkable passage, Karl R. Popper confesses very plainly to the impossibility of making a science out of only strictly verifiable and justifiable elements. (Holton, 1988, p. 41)

The idea that science can subject the totality of its presuppositions to experimental tests of truth or falsity has been roundly discredited for decades. Although no consensus has emerged as to what this means for a methodical logic of science (Gadamer, 1981, 1989; Nielsen & Lynch, 2022; Weinsheimer, 1985), there are certainly strong indications that an emphasis on the playful absorption into dialogical relationships offers a number of advantages for structuring a new perspective on the mutual implication of subject and object (Dawson et al., 2006; Fisher, 2004; Nersessian, 1996; Overton, 2002). Perhaps it is not unreasonable to begin systematically investigating other options offering alternatives to modern dualist assumptions of alienated subjects and objects (Fisher, 2019).

A great many cultural, economic, social, and psychological factors (Dewey, 1929; Faber, 1999a/b; Gigerenzer, 1993; Haraway, 2022; Kauffman & Roli, 2023; Kline, 1980; Overton, 2002; Prigogine, 1986, 1997; Prigogine & Stengers, 1984) contribute to the reasons why a new scientific paradigm has yet to cohere. But the time is past for clinging to counterproductive and obsolete ideas and methods, especially given the presence of viable alternatives. If quantitative methods merely accept naturally given units, then there would be no historical variation in the definitions of metric units or of the physical and social constructs worthy of investigation. Duncan then offers the relevant observation that:

> All measurement is . . . social measurement. Physical measures are made for social purposes and physical dimensions may be used by . . . scientists [in any domain of research or practice]. But social measurement in a narrower sense deals with phenomena that are beyond the ken of physics. To extend historical metrology to include social measurement, therefore, will require some modification of thought patterns. For one thing, we shall have to overcome our tendency to think of social measurement or quantification as something external to the social system in the sense, say, that the tailor's tape measure is external to the customer's waist. On the contrary, I argue, the quantification is implicit – sometimes explicit, for an observer not blinded by methodological preconceptions – in the social process itself before any social scientist intrudes. (Duncan, 1984b, pp. 35–36)

The roots of social measurement in social processes can be traced from the ancient Greek origins of mathematical thinking in Plato's accounts of Socratic dialogue (Fisher, 1988, 1992, 2003a/b, 2004, 2010). Rigorous conceptions of qualitatively meaningful quantification based in irreducible complexity (Commons et al., 2014; Dawson et al., 2006; Dawson-Tunik et al., 2005; Fischer & Dawson, 2002; Overton, 1998, 2002) can be seen to extend everyday language into mathematical scientific language (Fisher, 2019, 2020, 2021, 2023).

The social processes leading to the production of this volume involved a personal communication about Duncan between one of the authors (Fisher) and Benjamin Wright in the late 1980s. Wright stated that he had been unable to persuade Duncan to adopt Rasch's models during in their early years as young professors at the University of Chicago in the 1960s. Duncan later, however, grasped the essential differences between scientific and statistical models, worked through his understanding of the models from the bottom up, and forcefully advocated the adoption of Rasch's perspective in quantitative sociology (Duncan, 1984a/b/c, 1992; Duncan & Stenbeck, 1987, 1988).

But contrary to the arguments made by others as to the value of Rasch's ideas, Duncan did not frame his measurement perspective in terms of the choices made among models for data analyses. Instead, he was cognizant of the challenges and opportunities posed by metrological unit definitions:

> With the possible and, in any event, limited exception of economics, we have in social science no system of measurements that can be coherently described in terms of a small number of dimensions. Like physical scientists, we have thousands of 'instruments,' but these instruments purport to yield measurements of thousands of variables. That is, we have no system of units (much less standards for them) that, at least in principle, relates all of the variables to a common set of logically primitive qualities. There are no counterparts of mass, length and time in social science . . . . To the physical dimensions, economics adds money . . . . The fact that social science (beyond economics) does not have such a system of measurements is, perhaps, another way of saying that theory in our field is fragmentary and undeveloped, and that our knowledge is largely correlational rather than theoretical.

Significant advances in the development of predictive theories and explanatory models of measured constructs have occurred in the years before and since Duncan wrote (e.g., see Commons et al., 2014; De Boeck & Wilson, 2004; Embretson, 2010; Fischer, 1973; Green & Smith, 1987; Smith, 1996; Stenner & Smith, 1982; Stenner et al., 2013, 2016, 2023). Consistently reproducible correspondences of theory and evidence may be key factors substantiating a basis for confidence in systems of measurements traceable to a new class of candidate SI units. Documented instances (Barney, 2013, 2016; Barney & Fisher, 2016; Dawson, 2002, 2004; He, 2022; He & Kingsbury, 2016; Kingsbury, 2009; Pendrill 2019, 2024; Melin et al., 2021; Stenner et al., 2013; Stenner & Fisher, 2013; Williamson, 2018) of results demonstrating repeatable reproducibility of empirically stable and theoretically explained unit definitions set the stage for imagining, designing, and developing the kind of unit system Duncan has in mind. A major goal for us in compiling this book is simply to put this idea on the table as a serious matter for consideration.

The present volume, then, joins Duncan and many others (Cohen, 1994; Guttman, 1977, 1985; Meehl, 1967; Michell, 1986; Rodgers, 2010; Rogosa, 1987; Wilson, 2013a) in criticizing and offering alternatives to:

> the syndrome that I [Duncan] have come to call *statisticism*: the notion that computing is synonymous with doing research, the naive faith that statistics is a complete or sufficient basis for scientific methodology, the superstition that statistical formulas exist for evaluating such things as the relative merits of different substantive theories or the 'importance' of the causes of a 'dependent variable'; and the delusion that decomposing the covariations of some arbitrary and haphazardly assembled collection of variables can somehow justify not only a 'causal model' but also, praise the mark, a 'measurement model.' There would be no point in deploring such caricatures of the scientific enterprise if there were a clearly identifiable sector of social science research wherein such fallacies were clearly recognized and emphatically out of bounds. But in my discipline it just is not so. Individual articles of exemplary quality are published cheek-by-jowl with transpar-

> ent exercises in statistical numerology. If the muck were ankle deep, we could wade through it. When it is at hip level, our most adroit and most fastidious workers can hardly avoid getting some of it on their product. (Duncan, 1984b, pp. 226–227)

Focused new developments in the direction away from “statisticism” and toward consideration of the possible viability of extending the SI began in 2008. The record of events recounted above prompted initiatives aimed at possible collaborations with metrologists interested in transdisciplinary conceptual and operational overlaps with psychology and the social sciences.

1.4.3 New alliances

Upon investigation, it turned out that some metrologists (Beges et al., 2010; Berglund et al., 2012; Finkelstein, 1975, 1994, 2003, 2005; Mari, 2000, 2003, 2009; Mari et al., 2009; Mari & Sartori, 2007; Pendrill, 2008; Pendrill et al., 2010) had been seeking thought partners in psychology and the social sciences for years but had not identified additive conjoint models as a focal interest.

The International Measurement Confederation (IMEKO), whose membership is composed of globally distributed national metrology institutes, became the forum in which new alliances were formed. Conversations on measurement in physics and psychology ensued at Joint Symposia organized by the IMEKO Technical Committees on Measurement Science (TC7), Education and Training in Measurement and Instrumentation (TC1), and Measurements in Biology and Medicine (TC13), with the later addition of Measurements of Human Functions (TC18) (Fisher, 2008, 2010, 2012, 2014). At the 2008 IMEKO TC1-TC7-TC13 Joint Symposium held in Annecy, France, Ludwik Finkelstein opened a session (Finkelstein, 2008, 2009), introducing the next presentation (Fisher, 2009) and taking the opportunity to remark on a wide range of issues in the history and philosophy of science and measurement. Finkelstein noted that, in comparison with metrologists in the natural sciences, psychological measurement researchers had focused on model-based approaches to measurement on a broad scale for a longer time. He also expressed his personal opinion that, because natural scientists had the advantage of more thoroughly worked out theories, tools, methods, and standards, psychometricians in general were confronted with more difficult conceptual challenges than metrologists. He said that the formulation of measurement models setting forth clear inferential requirements for estimating interval quantities from ordinal observations was an important step forward in the advancement of measurement theory and practice, and that psychologists had made significant contributions deserving of increased attention in the natural sciences (Fisher, 2008).

At the 2010 IMEKO Joint Symposium hosted by Finkelstein, Sanowar Khan, Kenneth Grattan, and their colleagues in London, Finkelstein (2010, p. 2) said:

> The development of measurement science as a discipline has not paid adequate attention to the wider use of measurement. It is increasingly recognized that the wide range and diverse applications of measurement are based on common logical and philosophical principles and share common problems. However the concepts, vocabularies and methodologies in the various fields of measurement in the literature tend to differ. The development of a unified science of measurement appropriate for all domains of application seems to be desirable.

At that 2010 Joint Symposium, Luca Mari raised the question as to what it could mean for psychological and social instruments to be calibrated when calibration is always to a standard SI unit, which psychology and the social sciences do not have. That question immediately illuminated the nature of a mutually informative dialogue between the natural and social sciences (Fisher, 2010, p. 1279). On the one hand, the natural sciences lack, while psychology and the social sciences possess, decades of widely adopted traditions and norms concerning how ordinal observations can serve as a necessary and sufficient basis for estimating interval unit quantities and uncertainties. On the other hand, psychology and the social sciences lack the natural sciences' methods and expectations as to the value for communications, innovation, and commerce that stand to be obtained from distributed systems of instruments calibrated and metrologically traceable to quality assured unit standards.

Mari then suggested that a leader in psychological measurement modeling should be invited to give a special talk at the 2011 IMEKO Joint Symposium to be held in Jena, Germany (Scharff & Linß, 2011; Fisher, 2012a). Mark Wilson (2011, 2013b) gave that presentation and was backed up by several colleagues also presenting social and psychological measurement research applying Rasch's additive conjoint models (Bezruczko, 2011; Cano et al., 2011; Cooper & Fisher, 2011; Fisher, 2011; Fisher & Stenner, 2011; Granger & Bezruczko, 2011; Salzberger, 2011; Stenner et al., 2023). Several of these presentations were given in a session on fundamentals of measurement science chaired by Klaus-Dieter Sommer, one of the editors of the De Gruyter Series on Measurement Science in which this present volume appears.

Similar arrays of presentations on psychological and social measurement were given at subsequent IMEKO Joint Symposia in Genoa, Italy, in 2013, in Madeira in 2014, at the IMEKO World Congress in Prague in 2015, and at the 2016 Joint Symposium held at the University of California, Berkeley. This latter meeting was hosted by Wilson and Fisher (2016, 2018) and was the first such meeting attended by approximately equal numbers of social and natural scientists. At the following IMEKO World Congress in Belfast in 2018, a special session on psychological measurement was organized by Wilson and Fisher (2019), and significant participation by psychologists and social scientists continued at the 2017 Joint Symposium in Rio de Janeiro (Costa-Monteiro et al., 2018), Brazil; in St. Petersburg, Russia, in 2019 (Sapozhnikov & Taymanov, 2019); and in 2022 in Porto, Portugal (Benoit, 2022, 2023). Over the course of these recent years, Mari has productively collaborated with Wilson (Mari & Wilson, 2013, 2014) and Wilson's former students David Torres Irribarra and Andrew Maul (Mari et al., 2013, 2016, 2023; Maul et al., 2018, 2019; Wilson et al., 2015).

The first decade of the new millennium saw the emergence of MINET, a network audaciously entitled *Measuring the Impossible*, sponsored by the European Commission (Berglund et al., 2012; Pendrill, 2014; Pendrill et al., 2010). MINET brought together a multidisciplinary consortium of metrologists, physicists, engineers, psychophysicists, psychologists, and sociologists to address the challenges of measurements with persons from a metrological perspective. Keynote MINET researchers ranged from the BIPM Director Andrew Wallard to academics such as Giovanni Rossi (Genoa) and Birgitta Berglund (Karolinska) to well-known figures such as Ludwik Finkelstein, Fred Roberts, Damir Dzhafazov, and James Townsend. Parts of the MINET community subsequently merged with others, including IMEKO and IOMW (International Objective Measurement Workshop; Wright, 1992a, Wilson, 1992, 1994b; Engelhard & Wilson, 1996; Wilson et al., 1997; Wilson & Engelhard, 2000; Garner et al., 2010; Brown et al., 2011; Duckor et al., 2015). These associations have formed the basis of several collaborations (Cano et al., 2016, 2018a/b, 2019; Locoro et al., 2021; Melin et al., 2021; Pendrill & Fisher, 2013, 2015; Fisher et al., 2019), which have continued to bear fruit, as is evident in the production of several books in the *Springer Series in Measurement Science and Technology* (Fisher & Cano, 2023; Mari et al., 2023; Pendrill, 2019; Wilson & Fisher, 2017) as well as this volume.

1.5 Closing comments

Jeckelmann and Edelmaier (2023, p. 3) observe that "An extension of the concept of measurement will be necessary in future developments if the SI is to truly live up to its claim to be the universal language for all sciences." That extension need not, however, involve any further expansion of the oxymoronic concept of "ordinal quantity" or distinctions between "kinds of quantities" introduced in recent editions of the International Vocabulary of Measurement (JCGM, 2012, p. 15; Mari, 2009; Pendrill 2019). On the contrary, as was suggested by Finkelstein in his 2008 IMEKO talk in Annecy, it may be that the ongoing acceptance of ordinal scales for physical constructs such as hardness will soon give way to new representations in interval and ratio units. But an extended SI metrology for psychology and the social sciences will not be a mere elevation of those fields to a status akin to that of physics and the natural sciences. No, it would rather seem that exciting implications for a new art and science of complexity spanning the full range of fields are in store.

References

Ackermann, J. R. (1985). *Data, instruments, and theory: A dialectical approach to understanding science*. Princeton University Press.

Adams, R. J., Wu, M. L., Cloney, D., Berezner, A., & Wilson, M. (2020). ACER ConQuest: Generalised Item Response Modelling Software (Version 5.29). Australian Council for Educational Research. https://www.acer.org/au/conquest

Alder, K. (2002). *The measure of all things: The seven-year odyssey and hidden error that transformed the world*. The Free Press.

Allen, D. D., & Pak, S. (2023). Improving clinical practice with person-centered outcome measurement. In W. P. Fisher Jr. & S. J. Cano (Eds.). *Person centered outcome metrology* (pp. 53–105). Springer.

Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, *65*(331), 1248–1255.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*(1), 69–81.

Andersen, E. B. (1980). *Discrete statistical models with social science applications*. North-Holland.

Andersen, E. B. (1999). Sufficient statistics in educational measurement. In G. N. Masters & J. P. Keeves (Eds.). *Advances in measurement in educational research and assessment* (pp. 122–125). Pergamon.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

Andrich, D. (1988). *Sage University Paper Series on Quantitative Applications in the Social Sciences. Vol. series no. 07–068: Rasch models for measurement*. Sage Publications.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J. A. Keats, R. Taft, R. A. Heath, & S. H. Lovibond (Eds.), *Mathematical and Theoretical Systems: Proceedings of the 24th International Congress of Psychology of the International Union of Psychological Science, Vol. 4* (pp. 7–16). Elsevier Science Publishers.

Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika*, *75*(2), 292–308.

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social, and health sciences*. Springer.

Andrich, D., Sheridan, B., & Luo, G. (2017). *RUMM 2030: Rasch unidimensional models for measurement*. RUMM Laboratory Pty Ltd [www.rummlab.com.au].

Aryadoust, S. V., Tan, H. A. H., & Ng, L. Y. (2019). A scientometric review of Rasch measurement: The rise and progress of a specialty. *Frontiers in Psychology*, *10*, 2197.

Ashworth, W. J. (2004). Metrology and the state: Science, revenue, and commerce. *Science*, *306*(5700), 1314–1317.

Barney, M. (2013). *Leading value creation: Organizational science, bioinspiration, and the cue see model*. Palgrave Macmillan.

Barney, M. (2016). Calibrating charisma: The many-facet Rasch model for leader measurement and automated coaching. *Journal of Physics: Conference Series*, *772*(012051), doi: 10.1088/1742-6596/772/1/012051

Barney, M., & Barney, F. (2024). Transdisciplinary measurement through AI: Hybrid metrology and psychometrics powered by Large Language Models. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*. De Gruyter.

Barney, M., & Fisher, W. P., Jr. (2016). Adaptive measurement and assessment. *Annual Review of Organizational Psychology and Organizational Behavior*, *3*, 469–490.

Beges, G., Drnovsek, J., & Pendrill, L. R. (2010). Optimising calibration and measurement capabilities in terms of economics in conformity assessment. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement*, *15*(3), 147–154.

Benoit, E. (2022). Editorial to selected papers from IMEKO TC1-TC7-TC13-TC18 Joint Symposium and MATHMET (European Metrology Network for Mathematics and Statistics) Workshops 2022. *Acta IMEKO*, *11*(4), 1–2.

Benoit, E. (2023). Editorial to selected papers from IMEKO TC1-TC7-TC13-TC18 joint symposium and MATHMET workshop 2022, 2nd part. *Acta IMEKO*, *12*(2), 1–2.

Berglund, B., Rossi, G. B., Townsend, J. T., & Pendrill, L. R. (Eds.). (2012). *Measurement with persons: Theory, methods, and implementation areas*. Psychology Press, Taylor & Francis Group.

Bezruczko, N. (Ed.). (2005). *Rasch measurement in health sciences*. JAM Press.

Bezruczko, N. (2011, September 2). Foundational imperatives for measurement with mathematical models. K.-D. Sommer (Chair), *Session on Fundamentals of measurement science*. In P. Scharff & G. Linß (Eds.), *Proceedings of the 14th Joint International IMEKO TC1 + TC7 + TC 13 Symposium: Intelligent quality measurements – theory, education and training. 31 August – 2 September.* Ilmenau University of Technology. https://www.db-thueringen.de/receive/dbt_mods_00019432

Black, M. (1946). *Critical thinking*. Prentice Hall.

Black, M. (1962). *Models and metaphors*. Cornell University Press.

Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives*, *9*, 1–52.

Blok, A., Farias, I., & Roberts, C. (Eds.). (2020). *The Routledge companion to Actor-Network Theory*. Routledge.

Blok, A., Nakazora, M., & Winthereik, B. R. (2016). Infrastructuring environments. *Science as Culture*, *25*(1), 1–22.

Bode, R. K., Heinemann, A. W., & Semik, P. (2000). Measurement properties of the Galveston Orientation and Amnesia Test (GOAT) and improvement patterns during inpatient rehabilitation. *Journal of Head Trauma Rehabilitation*, *15*(1), 637–655.

Bohr, N. (1963). *Essays 1958–1962 on atomic physics and human knowledge*. John Wiley & Sons.

Boone, W. J., & Staver, J. R. (2020). *Advances in Rasch analyses in the human sciences*. Springer.

Bowker, G. C. (2016). How knowledge infrastructures learn. In P. Harvey, C. B. Jensen, & A. Morita (Eds.). *Infrastructures and social complexity: A companion* (pp. 391–403). Routledge.

Bowker, G., Timmermans, S., Clarke, A. E., & Balka, E. (Eds.) (2015). *Boundary objects and beyond: Working with Leigh Star*. MIT Press.

Brogden, H. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika*, *42*, 631–634.

Brown, N., Duckor, B., Draney, K., & Wilson, M. (Eds.). (2011). *Advances in Rasch measurement, Vol. Two*. JAM Press.

Bryman, A. (2007). Barriers to integrating quantitative and qualitative research. *Journal of Mixed Methods Research*, *1*(1), 8–22.

Bud, R., & Cozzens, S. E. (Eds.). (1992). *SPIE Institutes: Vol. 9. Invisible connections: Instruments, institutions, and science*. (R. F. Potter, Ed.). SPIE Optical Engineering Press.

Bulut, O. (2021). *Package ‘eirm’: Explanatory Item Response Modeling for Dichotomous and Polytomous Items*. University of Alberta, Edmonton: CRAN, https://github.com/okanbulut/eirm

Calude, C. S. (2002). Incompleteness, complexity, randomness and beyond. *Minds and Machines*, *12*, 503–517.

Calude, C. S. (Ed.). (2007). *Randomness & complexity from Leibniz to Chaitin*. World Scientific.

Cano, S., Klassen, A. F., & Pusic, A. L. (2011, September 2). *From Breast-Q © to Q-Score ©: Using Rasch measurement to better capture breast surgery outcomes*. In P. Scharff & G. Linß (Eds.), *Proceedings of the 14th Joint International IMEKO TC1 + TC7 + TC 13 Symposium: Intelligent quality measurements – theory, education and training. 31 August – 2 September.* Ilmenau University of Technology. http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24429/ilm1-2011imeko-039.pdf

Cano, S., Melin, J., Fisher, W. P., Jr., Stenner, A. J., & Pendrill, L., EMPIR NeuroMet 15HLT04 Consortium. (2018a). Patient-centred cognition metrology. *Journal of Physics: Conference Series*, *1065*, 072033. https://iopscience.iop.org/article/10.1088/1742-6596/1065/7/072033/meta

Cano, S., Pendrill, L., Barbic, S., & Fisher, W. P., Jr. (2018b). Patient-centred outcome metrology for healthcare decision-making. *Journal of Physics: Conference Series*, *1044*, 012057. http://iopscience.iop.org/article/10.1088/1742-6596/1044/1/012057

Cano, S., Pendrill, L., Melin, J., & Fisher, W. P., Jr. (2019). Towards consensus measurement standards for patient-centered outcomes. *Measurement*, *141*, 62–69. https://doi.org/10.1016/j.measurement.2019.03.056

Cano, S., Vosk, T., Pendrill, L., & Stenner, A. J. (2016). On trial: The compatibility of measurement in the physical and social sciences. *Journal of Physics: Conference Series*, *772*, 012025. doi: 10.1088/1742-6596/772/1/012025

Chaitin, G. J. (1994). Randomness and complexity in pure mathematics. *International Journal of Bifurcation and Chaos*, *4*(1), 3–15.

Chien, T. W., Chang, Y., Wen, K. S., & Uen, Y. H. (2018). Using graphical representations to enhance the quality-of-care for colorectal cancer patients. *European Journal of Cancer Care*, *27*(1), e12591.

Chien, T.-W., Linacre, J. M., & Wang, W.-C. (2011). Examining student ability using KIDMAP fit statistics of Rasch analysis in Excel. In H. Tan & M. Zhou (Eds.), *Communications in Computer and Information Science: Vol. 201. Advances in Information Technology and Education, CSE 2011 Qingdao, China Proceedings, Part I* (pp. 578–585). Springer Verlag.

Choi, E. (1998). Rasch invents “ounces.” *Popular Measurement*, *1*(1), 29. https://www.rasch.org/pm/pm1-29.pdf

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*, 186–190.

Cliff, N., & Keats, J. A. (2003). *Ordinal measurement in the behavioral sciences*. Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round (p < 0.05). *American Psychologist*, *49*, 997–1003. https://psycnet.apa.org/record/1995-12080-001

Commons, M. L., Gane-McCalla, R., Barker, C. D., & Li, E. Y. (2014). The model of hierarchical complexity as a measurement system. *Behavioral Development Bulletin*, *19*(3), 9–14.

Connolly, A. J., Nachtman, W., & Pritchett, E. M. (1971). *Keymath: Diagnostic Arithmetic Test*. American Guidance Service, https://images.pearsonclinical.com/images/pa/products/keymath3_da/km3-da-pub-summary.pdf

Cooper, G., & Fisher, W. P., Jr. (2011, September 2). *Continuous quantity and unit; their centrality to measurement*. In P. Scharff & G. Linß (Eds.), *Proceedings of the 14th Joint International IMEKO TC1 + TC7 + TC 13 Symposium: Intelligent quality measurements – theory, education and training. 31 August – 2 September.* Ilmenau University of Technology. http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24494/ilm1-2011imeko-019.pdf.

Costa-Monteiro, E., Costa-Felix, R. P. B., & Barbosa, C. R. H. (2018). 2017 Joint IMEKO TC1-TC7-TC13 Symposium: “Measurement Science Challenges in Natural and Social Sciences”. *Journal of Physics Conference Series, 1044,* (011001).

Crosby, A. W. (1997). *The measure of reality: Quantification and Western society, 1250-1600*. Cambridge University Press.

Dawson, T. L. (2002). New tools, new insights: Kohlberg’s moral reasoning stages revisited. *International Journal of Behavioral Development*, *26*(2), 154–166.

Dawson, T. L. (2004). Assessing intellectual development: Three approaches, one sequence. *Journal of Adult Development*, *11*(2), 71–85.

Dawson, T. L., Fischer, K. W., & Stein, Z. (2006). Reconsidering qualitative and quantitative research approaches: A cognitive developmental perspective. *New Ideas in Psychology*, *24*, 229–239.

Dawson-Tunik, T. L., Commons, M., Wilson, M., & Fischer, K. (2005). The shape of development. *The European Journal of Developmental Psychology*, *2*, 163–196.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. (Statistics for Social and Behavioral Sciences). Springer-Verlag.

de Courtenay, N. (2015). The double interpretation of the equations of physics and the quest for common meanings. In O. Schlaudt & L. Huber *Standardization in Measurement* (pp. 53–68), London: Pickering & Chatto Publishers.

Dewey, J. (1929). *The quest for certainty: A study of the relation of knowledge and action*. Perigee Books, G. P. Putnam's Sons.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model with the lme4 package. *Journal of Statistical Software*, *20*(2), http://www.jstatsoft.org/v20/i02

Duckor, B., & Holmberg, C. (2019). Exploring how to model formative assessment trajectories of posing-pausing-probing practices: Toward a teacher learning progressions framework for the study of novice teachers. *Journal of Educational Measurement*, *56*(4), 836–890.

Duckor, B., Santelices, M. V., & Brandt, S. (2015). Advances in Rasch modeling: New applications and directions for objective measurement science. *Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana*, *52*(2), 1–5. 10.7764/PEL.52.2.2015

Duncan, O. D. (1984a). Measurement and structure: Strategies for the design and analysis of subjective survey data. In C. F. Turner & E. Martin (Eds.). *Surveying subjective phenomena, Vol. 1* (pp. 179–229). Russell Sage Foundation.

Duncan, O. D. (1984b). *Notes on social measurement: Historical and critical*. Russell Sage Foundation.

Duncan, O. D. (1984c). Rasch measurement: Further examples and discussion. In C. F. Turner & E. Martin (Eds.). *Surveying subjective phenomena, Vol. 2* (pp. 367–403). Russell Sage Foundation.

Duncan, O. D. (1992). What if? *Contemporary Sociology*, *21*(5), 667–668.

Duncan, O. D., & Stenbeck, M. (1987). Are Likert scales unidimensional? *Social Science Research*, *16*(3), 245–259.

Duncan, O. D., & Stenbeck, M. (1988). Panels and cohorts: Design and model in the study of voting turnout. In C. C. Clogg (Ed.). *Sociological Methodology 1988* (pp. 1–35). American Sociological Association.

Edelmann, N. (2022). Digitalisation and developing a participatory culture: Participation, co-production, co-destruction. In Y. Charalabidis, L. S. Flak, & G. V. Pereira (Eds.). *Scientific foundations of digital governance and transformation (Public Administration and Information Technology, Vol. 38)* (pp. 415–435). Springer, https://doi.org/10.1007/978-3-030-92945-9_16

Einstein, A. (1954). *Ideas and opinions*. Bonanza Books.

Embretson, S. E. (1996a). Item Response Theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*(3), 201–212.

Embretson, S. E. (1996b). The new rules of measurement. *American Psychologist*, *8*(4), 341–349.

Embretson, S. E. (2010). *Measuring psychological constructs: Advances in model-based approaches*. American Psychological Association.

Engelhard, G., Jr. (2012). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge Academic.

Engelhard, G., Jr. & Wilson, M. (1996). *Objective measurement: Theory into practice, Vol. 3*. Ablex.

Faber, M. (1999a). Numbers and psychoanalysis: Reflections on the quest for certainty, Part I of II. *Psychoanalytic Review*, *86*(1), 63–107. https://www.proquest.com/scholarly-journals/numbers-psychoanalysis-reflections-on-quest/docview/195060395/se-2

Faber, M. (1999b). Numbers and psychoanalysis: Reflections on the quest for certainty, Part II of II. *Psychoanalytic Review*, *86*(2), 245–279. https://www.proquest.com/docview/195066081/citation/2E9355EFE0154496PQ/1?accountid=135705

Finkelstein, L. (1975). Representation by symbol systems as an extension of the concept of measurement. *Kybernetes*, *4*(4), 215–223.

Finkelstein, L. (1994). Measurement and instrumentation science: An analytical review. *Measurement*, *14*(1), 3–14.

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement*, *34*(1), 39–40 (10).

Finkelstein, L. (2005). Problems of measurement in soft systems. *Measurement*, *38*(4), 267–274.

Finkelstein, L. (2008, September 5). *Problems of widely-defined measurement*. Presented at the International Measurement Confederation (IMEKO) TC1 & TC7 Joint Symposium on Man, Science, & Measurement, Annecy, France.

Finkelstein, L. (2009). Widely-defined measurement–An analysis of challenges. *Measurement: Concerning Foundational Concepts of Measurement Special Issue Section*, *42*(9), 1270–1277.

Finkelstein, L. (2010). Measurement and instrumentation science and technology—the educational challenges. *Journal of Physics Conference Series*, *238*, 012001. doi: 10.1088/1742-6596/238/1/012001

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, *46*(1), 59–77.

Fischer, G. H., & Molenaar, I. (1995). *Rasch models: Foundations, recent developments, and applications*. Springer-Verlag.

Fischer, K. W., & Dawson, T. L. (2002). A new kind of developmental science: Using models to integrate theory and research: Comment. *Monographs of the Society for Research in Child Development*, *67*(1), 156–167.

Fisher, R. A. (1935). The logic of inductive inference [with discussion]. *Journal of the Royal Statistical Society*, *98*(1), 39–82. https://doi.org/10.2307/2342435

Fisher, W. P., Jr. (1988). Truth, method, and measurement: The hermeneutic of instrumentation and the Rasch model [Diss]. *Dissertation Abstracts International (*Dept. of Education, Division of the Social Sciences, University of Chicago*), 49*, 0778A. (376 pages, 23 figures, 31 tables)

Fisher, W. P., Jr. (1992). Objectivity in measurement: A philosophical history of Rasch's separability theorem. In M. Wilson (Ed.). *Objective measurement: Theory into practice. Vol. I* (pp. 29–58). Ablex Publishing Corporation.

Fisher, W. P., Jr. (2003a). The mathematical metaphysics of measurement and metrology: Towards meaningful quantification in the human sciences. In A. Morales (Ed.). *Renascent pragmatism: Studies in law and social science* (pp. 118–153). Ashgate Publishing Co.

Fisher, W. P., Jr. (2003b). Mathematics, measurement, metaphor, metaphysics: Parts I & II. *Theory & Psychology*, *13*(6), 753–828.

Fisher, W. P., Jr. (2004). Meaning and method in the social sciences. *Human Studies: A Journal for Philosophy and the Social Sciences*, *27*(4), 429–454.

Fisher, W. P., Jr. (2008). Notes on IMEKO symposium. *Rasch Measurement Transactions*, *22*(1), 1147. http://www.rasch.org/rmt/rmt221.pdf

Fisher, W. P., Jr (2009). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement*, *42*(9), 1278–1287.

Fisher, W. P., Jr. (2010). Unifying the language of measurement. *Rasch Measurement Transactions*, *24*(2), 1278–1281. http://www.rasch.org/rmt/rmt242.pdf

Fisher, W. P., Jr. (2011, September 1). Measurement, metrology and the coordination of sociotechnical networks. K.-D. Sommer (Chair), *Session on Fundamentals of measurement science*. In P. Scharff & G. Linß (Eds.), *Proceedings of the 14th Joint International IMEKO TC1 + TC7 + TC 13 Symposium: Intelligent quality measurements – theory, education and training. 31 August – 2 September.* Ilmenau University of Technology, http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24491/ilm1-2011imeko-017.pdf, Jena, Germany.

Fisher, W. P., Jr. (2012). 2011 IMEKO conference proceedings available online. *Rasch Measurement Transactions*, *25*(4), 1349. http://www.rasch.org/rmt/rmt254.pdf

Fisher, W. P., Jr. (2014). IMEKO 2008-2013 video presentations available online. *Rasch Measurement Transactions*, *27*(4), 1440–1441. https://www.rasch.org/rmt/rmt274b.htm

Fisher, W. P., Jr. (2019). A nondualist social ethic: Fusing subject and object horizons in measurement. *TMQ–Techniques, Methodologies, and Quality (Special Issue, Health Metrology)*, *10*, 21–40. https://portal2.ipt.pt/media/manager.php?src=servico&cmd=file&target=m1_MTc2NDQ#page=21

Fisher, W. P., Jr. (2020). Contextualizing sustainable development metric standards: Imagining new entrepreneurial possibilities. *Sustainability*, *12*(9661), 1–22. https://doi.org/10.3390/su12229661

Fisher, W. P., Jr. (2021). Bateson and Wright on number and quantity: How to not separate thinking from its relational context. *Symmetry*, *13*(1415), https://doi.org/10.3390/sym13081415

Fisher, W. P., Jr. (2023). Measurement systems, brilliant results, and brilliant processes in healthcare: Untapped potentials of person-centered outcome metrology for cultivating trust. In W. P. Fisher Jr. & S. Cano (Eds.). *Person-centered outcome metrology: Principles and applications for high stakes decision making* (pp. 357–396). Springer, https://link.springer.com/book/10.1007/978-3-031-07465-3

Fisher, W. P., Jr., & Cano, S. (Eds.). (2023). *Person-centered outcome metrology: Principles and applications for high stakes decision making*. Springer Series in Measurement Science & Technology. Springer, https://link.springer.com/book/10.1007/978-3-031-07465-3

Fisher, W. P., Jr., Oon, E. P.-T., & Benson, S. (2021). Rethinking the role of educational assessment in classroom communities: How can design thinking address the problems of coherence and complexity? *Educational Design Research*, *5*(1), 1–33. doi: 10.15460/eder.5.1.1537

Fisher, W. P., Jr, Pendrill, L., Lips da Cruz, A., & Felin, A. (2019). Why metrology? Fair dealing and efficient markets for the United Nations' Sustainable Development Goals. *Journal of Physics: Conference Series*, *1379*(012023), doi: 10.1088/1742-6596/1379/1/012023

Fisher, W. P., Jr., & Stenner, A. J. (2011, September 2). A technology roadmap for intangible assets metrology. In P. Scharff & G. Linß (Eds.). *International Measurement Confederation (IMEKO) TC1-TC7-TC13 Joint Symposium*, Jena, Germany, http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24493/ilm1-2011imeko-018.pdf, (Rpt. in W. P. Fisher, Jr., and P. J. Massengill (Eds.). (2023). *Explanatory models, unit standards, and personalized learning in educational measurement* (pp. 179–198). Springer, https://link.springer.com/book/10.1007/978-981-19-3747-7)

Fisher, W. P., Jr., & Wright, B. D. (Eds.). (1994). Applications of probabilistic conjoint measurement. *International Journal of Educational Research*, Vol. 21(6), 557–664.

Floyd, J., & Kanamori, A. (2016). Gödel vis-à-vis Russell: Logic and set theory to philosophy. In G. Crocco & E.-M. Engelen (Eds.). *Godelian Studies on the Max-Phil Notebooks, Vol. I* (pp. 243–326). Presses Universitaires de Provence.

Gadamer, H.-G. (1981). *Studies in Contemporary German Social Thought. Vol. 2: Reason in the age of science*. T. McCarthy (Ed.) (F. G. Lawrence, Trans.), MIT Press.

Gadamer, H.-G. (1989). *Truth and method*. (J. Weinsheimer & D. G. Marshall, Trans.) (Rev. ed.). Crossroad.

Galison, P. (1997). *Image and logic: A material culture of microphysics*. University of Chicago Press.

Galison, P. (1999). Trading zone: Coordinating action and belief. In M. Biagioli (Ed.). *The science studies reader* (pp. 137–160). Routledge.

Garfinkel, A. (1991). Reductionism. In R. Boyd, P. Gasper, & J. D. Trout (Eds.). *The philosophy of science* (pp. 443–459). MIT Press.

Garner, M., Engelhard, G., Jr., Fisher, W. P., Jr., & Wilson, M. (Eds.). (2010). *Advances in Rasch measurement, Volume One*. JAM Press.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.). *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Lawrence Erlbaum Associates.

Goodman, L. A. (2007). Otis Dudley Duncan, quantitative sociologist par excellence: Path analysis, loglinear methods, and Rasch models. *Research in Social Stratification and Mobility*, *25*(2), 129–139. doi: 10.1016/j.rssm.2007.05.005

Granger, C., & Bezruczko, N. (2011). Body, mind, and spirit are instrumental to functional health: A case study. In P. Scharff & G. Linß (Eds.), *Proceedings of the 14th Joint International IMEKO TC1 + TC7 + TC 13 Symposium: Intelligent quality measurements – theory, education and training. 31 August – 2 September.* Ilmenau University of Technology. https://www.db-thueringen.de/receive/dbt_mods_00019497

Green, K. E. (1986). Fundamental measurement: A review and application of additive conjoint measurement in educational testing. *Journal of Experimental Education*, *54*(3), 141–147.

Green, K. E., & Smith, R. M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics*, *12*(4), 369–381.

Guttman, L. (1977). What is not what in statistics. *The Statistician*, *26*, 81–107.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis*, *1*, 3–10.

Guttman, L. (1994). The mathematics in ordinary speech. In S. Levy (Ed.). *Louis Guttman on theory and methodology: Selected writings* (pp. 103–119). Dartmouth Publishing Company.

Hagell, P. (2014). Testing rating scale unidimensionality using the Principal Component Analysis (PCA)/t-test protocol with the Rasch model: The primacy of theory over statistics. *Open Journal of Statistics*, *4*(6), 456–465. doi: 10.4236/ojs.2014.46044

Hagell, P. (2019). Measuring Activities of Daily Living in Parkinson's disease: On a road to nowhere and back again? *Measurement*, *132*, 109–124. https://www.sciencedirect.com/science/article/pii/S0263224118308844

Haraway, D. J. (2022). A Cyborg Manifesto: An ironic dream of a common language for women in the integrated circuit. In *The Transgender Studies Reader Remix* (pp. 429–443). Routledge.

He, W. (2022). *MAP Growth item parameter drift study*. Portland, OR: NWEA Research Report.

He, W., Li, S., & Kingsbury, G. G. (2016). A large-scale, long-term study of scale drift: The micro view and the macro view. *Journal of Physics Conference Series*, *772*, 012022. https://iopscience.iop.org/article/10.1088/1742-6596/772/1/012022/meta

Hesse, M. (1970). *Models and analogies in science*. Notre Dame University Press.

Hohensinn, C. (2018). PcIRT: An R package for polytomous and continuous Rasch models. *Journal of Statistical Software*, *84*(2), 1–14. https://www.jstatsoft.org/article/view/v084c02/0

Holton, G. (1988). *Thematic origins of scientific thought: Kepler to Einstein* (revised ed.). Harvard University Press.

Huxley, T. H. (1862). *On our knowledge of the causes of the phenomena of organic nature*. Robert Hardwicke.

Ihde, D. (1991). *Instrumental realism: The interface between philosophy of science and philosophy of technology*. (The Indiana Series in the Philosophy of Technology). Indiana University Press.

Jasanoff, S. (2004). *States of knowledge: The co-production of science and social order*. (International Library of Sociology). Routledge.

JCGM: Joint Committee for Guides in Metrology. Working Group 2, (2012). *International vocabulary of metrology: Basic and general concepts and associated terms*, *3rd ed (with minor corrections)*, International Bureau of Weights and Measures–BIPM.

Jeckelmann, B., & Edelmaier, R. (Eds.). (2023). *Metrological infrastructure*. (K.-D. Sommer & T. Fröhlich, Eds.). (De Gruyter Series in Measurement Sciences). De Gruyter Oldenbourg, https://doi.org/10.1515/9783110715835

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449–1475.

Kahneman, D. (2011). *Thinking fast and slow*. Farrar, Straus and Giroux.

Karlsson, S., Svensson, H., Wisén, J., & Melin, J. (2024). A metrological approach to social sustainability metrics in municipalities. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 379–398), De Gruyter.

Kauffman, S. A., & Roli, A. (2023). A third transition in science? *Interface Focus*, *13*(3), 20220063.

Kielhofner, G., Dobria, L., Forsyth, K., & Basu, S. (2005). The construction of keyforms for obtaining instantaneous measures from the occupational performance history interview ratings scales. *OTJR: Occupation Participation and Health*, *25*(1), 23–32.

Kingsbury, G. G. (2009). Tools for measuring academic growth. *Journal of Applied Measurement*, *10*(1), 97.

Kline, M. (1980). *Mathematics: The loss of certainty*. Oxford University Press.

Knorr Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Harvard University Press.

Kollias, C. (2024). Placing multiple panel cut scores on the same measurement scale. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 345–360), De Gruyter.

Kula, W. (1986). *Measures and men*. (R. Screter, Trans.), Princeton University Press, (Original work published 1970).

Kumar, D. N., & Joshi, H. B. (2024). Patient-centred outcome assessments in the surgical disciplines. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 449–472), De Gruyter.

Lamprianou, I. (2020). *Applying the Rasch model in social sciences using R and BlueSky statistics*. Quantitative Methodology Series. Routledge.

Lang, W. S., & Wilkerson, J. (2024). Measuring teacher dispositions: Steps in an innovative journey in affective assessment. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 301–344), De Gruyter.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*. Sage.

Lerner, R. M., & Overton, W. F. (2017). Reduction to absurdity: Why epigenetics invalidates all models involving genetic reduction. *Human Development*, *60*, 107–123. https://doi.org/10.1159/000477995

Li, Y. (2006). Using the open-source statistical language R to analyze the dichotomous Rasch model. *Behavioral Research Methods*, *38*(3), 532–541.

Linacre, J. M. (1997). Instantaneous measurement and diagnosis. *Physical Medicine and Rehabilitation State of the Art Reviews*, *11*(2), 315–324. http://www.rasch.org/memo60.htm

Linacre, J. M. (2023a). Advancing the metrological agenda in the social sciences. In W. P. Fisher Jr. & S. Cano (Eds.). *Person-centered outcome metrology: Principles and applications for high stakes decision making* (pp. 165–193). Springer, https://link.springer.com/book/10.1007/978-3-031-07465-3

Linacre, J. M. (2023b). *A user's guide to FACETS Rasch-Model computer program, v. 3.85.1*. Winsteps.com, http://www.winsteps.com/a/facets-manual.pdf

Linacre, J. M. (2024). *A user's guide to WINSTEPS Rasch-Model computer program, v. 5.7.0*. Winsteps.com, https://www.winsteps.com/manuals.htm

Linacre, J. M., Chan, G., & Adams, R. J. (2013). An 'estimation bias' shootout in the wild West: CMLE, JMLE, MMLE, PMLE. *Rasch Measurement Transactions*, *27*(1), 1403–1405. http://www.rasch.org/rmt/rmt271.pdf

Liu, J. (2018). Development and translation of measurement findings for the motivation assessment for team readiness, integration, and collaboration self-scoring form. *American Journal of Occupational Therapy*, *72*(4_Supplement_1), 7211500015p1-7211500015p1.

Locoro, A., Fisher, W. P., Jr, & Mari, L. (2021). Visual information literacy: Definition, construct modeling and assessment. *IEEE Access*, *9*, 71053–71071. https://doi.org/10.1109/ACCESS.2021.3078429

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review*, *72*(2), 143–155.

Luce, R. D. (1959). *Individual choice behavior*. Wiley.

Luce, R. D. (1978). Dimensionally invariant numerical laws correspond to meaningful qualitative relations. *Philosophy of Science*, *45*, 1–16.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new kind of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27.

Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology*, *31*, 19–26.

MacIntosh, R. (1998). Global attitude measurement: An assessment of the World Values Survey Postmaterialism Scale. *American Sociological Review*, *63*(3), 452–464.

Mallinson, T. (2024). Extending the justice-oriented, anti-racist framework for validity testing to the application of measurement theory in re(developing) rehabilitation assessments. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 399–426). De Gruyter.

Mari, L. (2000). Beyond the representational viewpoint: A new formalization of measurement. *Measurement*, *27*, 71–84.

Mari, L. (2003). Epistemology of measurement. *Measurement*, *34*(1), 17–30.

Mari, L. (2009). On (kinds of) quantities. *Metrologia*, *46*(3), L11–L15. https://doi.org/10.1088/0026-1394/46/3/N01

Mari, L., Lazzarotti, V., & Manzini, R. (2009). Measurement in soft systems: Epistemological framework and a case study. *Measurement*, *42*(2), 241–253.

Mari, L., Maul, A., Torres Irribarra, D., & Wilson, M. (2013). Quantification is neither necessary nor sufficient for measurement. *Journal of Physics Conference Series*, *459*(1), http://iopscience.iop.org/1742-6596/459/1/012007

Mari, L., Maul, A., Torres Irribarra, D., & Wilson, M. (2016). Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement*, *100*, 115–121. http://www.sciencedirect.com/science/article/pii/S0263224116307497

Mari, L., & Sartori, L. (2007). A relational theory of measurement: Traceability as a solution to the non-transitivity of measurement results. *Measurement*, *40*, 233–242.

Mari, L., & Wilson, M. (2013). A gentle introduction to Rasch measurement models for metrologists [abstract]. *Journal of Physics Conference Series*, *459*(012002), http://iopscience.iop.org/1742-6596/459/1/012002/pdf/1742-6596_459_1_012002.pdf

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement*, *51*, 315–327. http://www.sciencedirect.com/science/article/pii/S0263224114000645

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement, 2nd ed*. (Springer Series in Measurement Science and Technology). Springer, https://link.springer.com/book/10.1007/978-3-031-22448-5

Marmor, Y. N., & Bashkansky, E. (2020). Processing new types of quality data. *Quality and Reliability Engineering International*, *36*, 2621–2638. https://doi.org/10.1002/qre.2642

Massof, R. W. (2008). Editorial: Moving toward scientific measurements of quality of life. *Ophthalmic Epidemiology*, *15*, 209–211.

Massof, R. W., Bradley, C., & McCarthy, A. M. (2024). Constructing a continuous latent disease state variable from clinical signs and symptoms. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 269–300), De Gruyter.

Masters, G. N., & Keeves, J. P. (Eds.). (1999). *Advances in measurement in educational research and assessment*. Pergamon.

Maul, A., Mari, L., Torres Irribarra, D., & Wilson, M. (2018). The quality of measurement results in terms of the structural features of the measurement process. *Measurement*, *116*, 611–620.

Maul, A., Mari, L., & Wilson, M. (2019). Intersubjectivity of measurement across the sciences. *Measurement*, *131*, 764–770. https://www.sciencedirect.com/science/article/pii/S026322411830808X?via%3Dihub

Maxwell, J. C. (1873). *A treatise on electricity and magnetism, Vol. 1*. Clarendon press.

Mead, R. J. (2009). The ISR: Intelligent Student Reports. *Journal of Applied Measurement*, *10*(2), 208–224.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103–115.

Melin, J. (2024). Is validity a straightforward concept to be used in measurements in the human and social sciences? In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 157–190). De Gruyter.

Melin, J., Cano, S. J., Flöel, A., Göschel, L., & Pendrill, L. (2021). Construct specification equations: 'recipes' for certified reference materials in cognitive measurement. *Measurement: Sensors*, *18*, 100290. https://doi.org/10.1016/j.measen.2021.100290

Melin, J., & Pendrill, L. R. (2023). The role of construct specification equations and entropy in the measurement of memory. In W. P. Fisher Jr & S. Cano (Eds.). *Person-centered outcome metrology: Principles and applications for high stakes decision making* (pp. 269–309). Springer Series in Measurement Science and Technology. Springer, https://link.springer.com/book/10.1007/978-3-031-07465-3

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, *100*, 398–407.

Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A construct-modeling approach to develop a learning progression of how students understand the structure of matter. *Journal of Research in Science Teaching*, *54*(8), 1024–1048.

Morell, L., Tan, S., & Wilson, M. (2024). Relationship between measures of "21st century skills" and the content underlying them. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 427–448). De Gruyter.

Moulton, M. (1993). Probabilistic mapping. *Rasch Measurement Transactions*, *7*(1), 268. http://www.rasch.org/rmt/rmt71b.htm

Nagel, E., & Newman, J. R. (1958). *Gödel's Proof*. New York University Press.

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, *99*(2), 166–180.

Nersessian, N. J. (1996). Child's play. *Philosophy of Science*, *63*, 542–546.

Nersessian, N. J. (2008). *Creating scientific concepts*. MIT Press.

Newby, V. A., Conner, G. R., Grant, C. P., & Bunderson, C. V. (2009). The Rasch model and additive conjoint measurement. *Journal of Applied Measurement*, *10*(4), 348–354.

Nielsen, C. R., & Lynch, G. (2022). *Gadamer's Truth and Method: A polyphonic commentary*. Rowman & Littlefield.

Overton, W. F. (1998). Developmental psychology: Philosophy, concepts, and methodology. *Handbook of Child Psychology*, *1*, 107–188.

Overton, W. F. (2002). Understanding, explanation, and reductionism: Finding a cure for Cartesian anxiety. In L. Smith & T. Brown (Eds.). *Reductionism and the development of knowledge* (pp. 29–51). Erlbaum.

Pelton, T., & Bunderson, V. (2003). The recovery of the density scale using a stochastic quasi-realization of additive conjoint measurement. *Journal of Applied Measurement*, *4*(3), 269–281.

Pendrill, L. R. (2008). Operating 'cost' characteristics in sampling by variable and attribute. *Accreditation and Quality Assurance*, *13*(11), 619–631.

Pendrill, L. R. (2014). Man as a measurement instrument [Special Feature]. *NCSLi Measure: The Journal of Measurement Science*, *9*(4), 22–33. http://www.tandfonline.com/doi/abs/10.1080/19315775.2014.11721702

Pendrill, L. R. (2019). *Quality assured measurement: Unification across social and physical sciences*. (Springer Series in Measurement Science and Technology). Springer, https://link.springer.com/book/10.1007/978-3-030-28695-8

Pendrill, L. R. (2024). Quantities and units: Order amongst complexity. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 35–100). De Gruyter.

Pendrill, L. R., Emardson, R., Berglund, B., Gröning, M., Höglund, A., Cancedda, A., Quinti, G., Crenna, F., Rossi, G. B., Drnovsek, J., Gersak, G., Van der Heijden, G., Kallinen, K., & Ravaja, N. (2010).

Measurement with persons: A European network. *NCSLi Measure: The Journal of Measurement Science*, *5*(2), 42–54. https://doi.org/10.1080/19315775.2010.11721515

Pendrill, L., & Fisher, W. P., Jr (2013). Quantifying human response: Linking metrological and psychometric characterisations of man as a measurement instrument. *Journal of Physics Conference Series*, *459*, http://iopscience.iop.org/1742-6596/459/1/012057

Pendrill, L., & Fisher, W. P., Jr (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement*, *71*, 46–55. doi:http://dx.doi.org/10.1016/j.measurement.2015.04.010

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*(2), 237–255.

Pesudovs, K. (2006). Patient-centred measurement in ophthalmology – A paradigm shift. *BMC Ophthalmology*, *6*(25).

Pesudovs, K. (2010). Item banking: A generational change in patient-reported outcome measurement. *Optometry and Vision Science*, *87*(4), 285–293.

Power, M. (2004). Counting, control, and calculation: Reflections on measuring and management. *Human Relations*, *57*(6), 765–783. doi:10.1177/0018726704044955

Prigogine, I. (1986). Science, civilization and democracy: Values, systems, structures and affinities. *Futures*, *18*(4), 493–507.

Prigogine, I. (1997). *The end of certainty: Time, chaos, and the new laws of nature*. Free Press.

Prigogine, I., & Stengers, I. (1984). *Order out of chaos: Man's new dialogue with nature*. Bantam Books.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Danmarks Paedogogiske Institut.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability: Volume IV: Contributions to biology and problems of medicine* (pp. 321–333 [http://www.rasch.org/memo1960.pdf]). University of California Press.

Rasch, G. (1966a). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.). *Readings in mathematical social science* (pp. 89–108). Science Research Associates, https://www.rasch.org/memo19662.pdf

Rasch, G. (1966b). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, *19*, 49–57.

Robinson, M., Johnson, A. M., Walton, D. M., & MacDermid, J. C. (2019). A comparison of the polytomous Rasch analysis output of RUMM2030 and R (ltm/eRm/TAM/lordif). *BMC Medical Research Methodology*, *19*(1), 36.

Robitzsch, A., Kiefer, T., & Wu, M., (2020). *TAM: Test Analysis Modules. R package version 3.5–19*. Salzburg, Austria: Federal Institute for Education Research, Innovation and Development of the Austrian School System, https://CRAN.R-project.org/package=TAM

Roche, J. (1998). *The mathematics of measurement: A critical history*. The Athlone Press.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*, 1–12. https://doi.org/10.1037/a0018326

Rogosa, D. (1987). Casual [sic] models do not support scientific conclusions: A comment in support of Freedman. *Journal of Educational Statistics*, *12*(2), 185–195.

Salzberger, T. (2009). *Measurement in marketing research: An alternative framework*. Edward Elgar.

Salzberger, T. (2011). The quantification of latent variables in the social sciences: Requirements for scientific measurement and shortcomings of current procedures. K.-D. Sommer (Chair), *Session on Fundamentals of measurement science*. In P. Scharff & G. Linß (Eds.), *Proceedings of the 14th Joint International IMEKO TC1 + TC7 + TC 13 Symposium: Intelligent quality measurements – theory, education* *and training. 31 August – 2 September*, *Jena, Germany.* Ilmenau University of Technology. https://www.db-thueringen.de/receive/dbt_mods_00019432

Sampson, G., Burgess, Y., McBride, N., & Steveno, B. (2024). A many-faceted measurement modeling approach for informing test specifications: Practical guidance from the national registry of emergency medical technicians. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 361–378). De Gruyter.

San Martin, E., Gonzalez, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3 PL model. *Psychometrika*, *80*(2), 450–467.

San Martin, E., & Rolin, J. M. (2013). Identification of parametric Rasch-type models. *Journal of Statistical Planning and Inference*, *143*(1), 116–130.

San Martin, E., Perticará, M., Varas, I. M., Kenzo Asahi, K., & González, J. (2024). The role of identifiability in empirical research. In W. P. Fisher Jr. & L. R. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 131–156). De Gruyter.

Sapozhnikova, K., & Taymanov, R. (2019). Joint IMEKO TC1-TC7-TC13-TC18 Symposium, 2–5 July, St. Petersburg, Russia. *Journal of Physics: Conference Series*, *1379*, 011001. doi: 10.1088/1742-6596/1379/1/011001

Scharff, P., & Linß, G. (Eds.). (2011). *Intelligent Quality Measurements – Theory*, *Education and Training: Proceedings of the 14th Joint International IMEKO TC1+TC7+TC13 Symposium*, *31 Aug – 2 Sept., Jena, Germany.* Ilmenau University of Technology (13 pp.), https://www.db-thueringen.de/receive/dbt_mods_00018047?q=imeko

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press.

Simon, H. A. (1997). *Models of bounded rationality: Empirically grounded economic reason Vol. 3*. MIT press.

Simon, H. A. (2000). Bounded rationality in social science: Today and tomorrow. *Mind & Society*, *1*, 25–39.

Smith, E. V., Jr, & Smith, R. M. (2004). *Introduction to Rasch measurement*. JAM Press.

Smith, E. V., Jr, & Smith, R. M. (2007). *Rasch measurement: Advanced and specialized applications*. JAM Press.

Smith, R. M. (1994). Reporting candidate performance on computer adaptive tests: IPARM. *Rasch Measurement Transactions*, *8*(1), 344–345.

Smith, R. M. (1996). Item component equating. In M. Wilson (Ed.). *Objective measurement: Theory into practice, Vol. 3* (pp. 289–308). Ablex Publishing Co.

Smith, R. M. (1997). The relationship between goals and functional status in the Patient Evaluation Conference System. *Physical Medicine and Rehabilitation: State of the Art Reviews*, *11*(2), 333–343.

Stenner, A. J., & Fisher, W. P., Jr (2013). Metrological traceability in the social sciences: A model from reading measurement. *Journal of Physics Conference Series*, *459*(012025), http://iopscience.iop.org/1742-6596/459/1/012025

Stenner, A. J., Fisher, W. P., Jr, Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology: Quantitative Psychology and Measurement*, *4*(536), 1–14. doi:, 10.3389/fpsyg.2013.00536. (Rpt. in W. P. Fisher, Jr. & P. J. Massengill (Eds.). (2023). *Explanatory models, unit standards, and personalized learning in educational measurement* (pp. 223–250). Springer, https://link.springer.com/book/10.1007/978-981-19-3747-7?page=1#toc)

Stenner, A. J., Fisher, W. P., Jr., Stone, M. H., & Burdick, D. S. (2016). Causal Rasch models in language testing: An application rich primer. In Q. Zhang (Ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings* (pp. 1–14). Springer.

Stenner, A. J., & Smith, M., III. (1982). Testing construct theories. *Perceptual and Motor Skills*, *55*, 415–426. (Rpt. in W. P. Fisher, Jr. & P. J. Massengill (Eds.). (2023). *Explanatory models, unit standards, and personalized learning in educational measurement* (pp. 31–42). Springer, https://link.springer.com/book/10.1007/978-981-19-3747-7?page=1#toc)

Stenner, A. J., Stone, M., & Burdick, D. (2023). How to model and test for the mechanisms that make measurement systems tick. In W. P. Fisher Jr. & P. J. Massengill (Eds.). *Explanatory models, unit* *standards, and personalized learning . . .* (pp. 199–211). Springer, https://link.springer.com/content/pdf/10.1007/978-981-19-3747-7_15?pdf=chapter%20toc, (Originally presented in session on Fundamentals of measurement science, K.-D. Sommer (Chair). International Measurement Confederation (IMEKO), Jena, Germany, 31 August 2011, https://www.db-thueringen.de/receive/dbt_mods_00019430)

Stephanou, A., & Fisher, W. P., Jr. (2013). From concrete to abstract in the measurement of length. *Journal of Physics Conference Series*, *459*, http://iopscience.iop.org/1742-6596/459/1/012026

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181.

Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman & P. Ratoosh (Eds.). *Measurement: Definitions and theories* (pp. 18–63). Wiley.

Sul, D. (2024). Situating culturally specific assessment development within the disjuncture-response dialectic. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*, (pp. 473–498). De Gruyter.

Tennant, A., Penta, M., Tesio, L., Grimby, G., Thonnard, J.-L., Slade, A., Lawton, G., Simone, A., Carter, J., Lundgren-Nilsson, A., Tripolski, M., Ring, H., Biering-Sorensen, F., Marincek, C., Burger, H., & Phillips, S. (2013). Assessing and adjusting for cross cultural validity of impairment and activity limitation scales through Differential Item Functioning within the framework of the Rasch model: The Pro-ESOR project. *Medical Care*, *42*(1), 137–148. http://www.jstor.org/stable/4640700

Teresi, J., Golden, R. R., Cross, P., Gurland, B., Kleinman, M., & Wilder, D. (1995). Item bias in cognitive screening measures: Comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. *Journal of Clinical Epidemiology*, *48*(4), 473–483.

Thurstone, L. L. (1959). Attitudes can be measured. In L. L. Thurstone, *The measurement of values* (pp. 215–233). University of Chicago Press, Midway Reprint Series. (Rpt. from Thurstone, L. L. (1928). *American Journal of Sociology*, *XXXIII*, 529–544.).

Torres Irribarra, D., & Freund, R. (2014). Wright map package for R. *Rasch Measurement Transactions*, *28*(1), 1456. http://www.rasch.org/rmt/rmt281.pdf

Toulmin, S. E. (1953). *The philosophy of science: An introduction*. Hutchinson's University Library.

Tukey, J. A. (1986). Data analysis and behavioural science. In L. V. Jones (Ed.), *The collected works of John A. Tukey, Volume III, Philosophy and principles of data analysis: 1949 – 1964* (pp. 187–390). Chapman & Hall.

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.). *Rasch models: Foundations recent developments, and applications* (pp. 215–237). Springer.

Verhelst, N., Hatzinger, R., & Mair, P. (2007). The Rasch sampler. *Journal of Statistical Software*, *20*(4), http://www.jstatsoft.org/v20/i04

von Davier, M., & Carstensen, C. H. (2007). *Multivariate and mixture distribution Rasch models: Extensions and applications*. Springer.

Weinsheimer, J. (1985). *Gadamer's hermeneutics: A reading of Truth and Method*. Yale University Press.

Weiss, D. J. (2021). *The Rasch model*. Retrieved 6 March 2024, from Assessment System Corp.: https://assess.com/the-rasch-model/.

Weitzel, T. (2004). *Economics of standards in information networks*. Physica-Verlag.

Williamson, G. L. (2018). Exploring reading and mathematics growth through psychometric innovations applied to longitudinal data. *Cogent Education*, *5*(1464424), 1–29.

Wilson, M. R. (1992). *Objective measurement: Theory into practice, Vol. 1*. Ablex.

Wilson, M. R. (1994a). Comparing attitude across different cultures: Two quantitative approaches to construct validity. In M. Wilson (Ed.). *Objective measurement: Theory into practice, Vol. 2* (pp. 271–294). Ablex.

Wilson, M. R. (1994b). *Objective measurement: Theory into practice, Vol. 2*. Ablex.

Wilson, M. R. (2011, September 1). The role of mathematical models in measurement: A perspective from psychometrics. L. Mari (Chair), *Plenary lecture*. In P. Scharff & G. Linß (Eds.), *Proceedings of the 14th Joint International IMEKO TC1 + TC7 + TC 13 Symposium: Intelligent quality measurements – theory, education and training. 31 August – 2 September, Jena, Germany.* http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24178/ilm1-2011imeko-005.pdf, Jena, Germany.

Wilson, M. R. (2013a). Seeking a balance between the statistical and scientific elements in psychometrics. *Psychometrika*, *78*(2), 211–236.

Wilson, M. R. (2013b). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement*, *46*, 3766–3774. http://www.sciencedirect.com/science/article/pii/S0263224113001061

Wilson, M. R. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice*, *37*(1), 5–20.

Wilson, M. R. (2023). *Constructing measures: An item response modeling approach, 2nd ed*. Routledge, https://doi.org/10.4324/9781410611697

Wilson, M., & Engelhard, G. (2000). *Objective measurement: Theory into practice*, Vol. 5. Ablex Publishing.

Wilson, M., Engelhard, G., & Draney, K. (Eds.). (1997). *Objective measurement: Theory into practice, Vol. 4*. Ablex.

Wilson, M., & Fisher, W. P., Jr. (2016). Preface: 2016 IMEKO TC1-TC7-TC13 Joint Symposium: Metrology across the Sciences: Wishful Thinking? *Journal of Physics Conference Series*, *772*(1), 011001. http://iopscience.iop.org/article/10.1088/1742-6596/772/1/011001/pdf

Wilson, M., & Fisher, W. P., Jr. (Eds.). (2017). *Psychological and social measurement: The career and contributions of Benjamin D. Wright*. (Springer Series in Measurement Science and Technology.) Springer Nature, https://link.springer.com/book/10.1007/978-3-319-67304-2

Wilson, M., & Fisher, W. P., Jr. (2018). Preface of special issue Metrology across the Sciences: Wishful Thinking? *Measurement*, *127*, 577. https://doi.org/10.1016/j.measurement.2018.07.028

Wilson, M., & Fisher, W. P., Jr. (2019). Preface of special issue, psychometric metrology. *Measurement*, *145*, 190. https://www.sciencedirect.com/journal/measurement/special-issue/10C49L3R8GT

Wilson, M., Mari, L., Maul, A., & Torres Irribarra, D. (2015). A comparison of measurement concepts across physical science and social science domains: Instrument design, calibration, and measurement. *Journal of Physics Conference Series*, *588*(012034), http://iopscience.iop.org/1742-6596/588/1/012034

Wilson, M., Scalise, K., & Gochyyev, P. (2019). Domain modelling for advanced learning environments: The BEAR assessment system software. *Educational Psychology*, *39*(10), 1199–1217.

Wise, M. N. (1995). Precision: Agent of unity and product of agreement. Part III–"Today Precision Must Be Commonplace." In M. N. Wise (Ed.). *The values of precision* (pp. 352–361). Princeton University Press.

Wittgenstein, L. (1983). *Remarks on the foundations of mathematics*. G. H. von Wright, R. Rhees, & G. E. M. Anscombe (Eds.). (G. E. M. Anscombe, Trans.), MIT Press.

Wolfe, F., Van der Heijde, D. M., & Larsen, A. (2000). Assessing radiographic status of rheumatoid arthritis: Introduction of a short erosion scale. *Journal of Rheumatology*, *27*(9), 2090–2099.

Wood, R. (1978). Fitting the Rasch model: A heady tale. *British Journal of Mathematical and Statistical Psychology*, *31*, 27–32.

Woolley, A. W., & Fuchs, E. (2011). Collective intelligence in the organization of science. *Organization Science*, *22*(5), 1359–1367.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 invitational conference on testing problems* (pp. 85–101). Educational Testing Service, http://www.rasch.org/memo1.htm

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97–116. http://www.rasch.org/memo42.htm

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review*, *3*(1), 281–288. http://www.rasch.org/memo41.htm

Wright, B. D. (1988). Georg Rasch and measurement: Informal remarks by Ben Wright at the Inaugural Meeting of the AERA Rasch Measurement SIG, New Orleans – April 8, 1988. *Rasch Measurement Transactions*, *2*, 25–32. http://www.rasch.org/rmt/rmt23.htm

Wright, B. D. (1992a). The International Objective Measurement Workshops: Past and future. In M. Wilson (Ed.). *Objective measurement: Theory into practice, Vol. 1* (pp. 9–28). Ablex Publishing.

Wright, B. D. (1992b). Raw scores are not linear measures. *Rasch Measurement Transactions*, *6*(1), 208. http://www.rasch.org/rmt/rmt61n.htm

Wright, B. D. (1997a). Fundamental measurement for outcome evaluation. *Physical Medicine & Rehabilitation State of the Art Reviews*, *11*(2), 261–288.

Wright, B. D. (1997b). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*(4), 33–45, 52. http://www.rasch.org/memo62.htm, https://doi.org/10.1111/j.1745-3992.1997.tb00606.x

Wright, B. D. (2012). Benjamin D. Wright's annotated KeyMath Diagnostic Profile. *Rasch Measurement Transactions*, *25*(4), 1350. https://www.rasch.org/rmt/rmt254.pdf

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*(12), 857–867. http://www.rasch.org/memo44.htm

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. MESA Press.

Wright, B. D., Mead, R. J., & Ludlow, L. H. (1980). *KIDMAP: person-by-item interaction mapping* (MESA Memorandum #29). Chicago: MESA Press [http://www.rasch.org/memo29.pdf]. (6 pp.)

Wright, B. D., & Stenner, A. J. (1998). Readability and reading ability. Paper Presented to the Australian Council on Education Research (ACER), Melbourne, Australia. (Rpt. in W. P. Fisher, Jr. & P. J. Massengill (Eds.). (2023). *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement* (pp. 89–107). Springer).

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. MESA Press.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials*. Wide Range, Inc, http://www.rasch.org/measess/me-all.pdf

Yumoto, F., & Stone, M. H. (2011). Comparing item calibration estimates using Winsteps and RUMM2010. *Rasch Measurement Transactions, 25*(3), 1337. https://www.rasch.org/rmt/rmt253e.htm

Part I: **Theory and Principles in Measurement and Metrology**

Leslie R. Pendrill

2 Quantities and units: order among complexity

Abstract: Modeling of concepts of quantities, units, and relations between them supports better inferences based on measurements made by end-users when tackling demanding issues in whatever context. A wider perspective is adopted, where the "entity" to which quantities and units are attributed can in principle be anything. Measurement quantities and units are in our view only a subset of quantities and units in general, albeit essential.

Models of quantities, units, and measurement systems, traditionally formed in the physical and engineering sciences, can also be sought in everyday language and even aesthetics. Scientific language can make implicit mathematical inferences explicit but does need to be extended to include categorical data, that is, as a result of classifications of measurement system response, even on the less-quantitative ordinal and nominal scales.

A recurring theme throughout the chapter is the relative proportions of order and complexity when modeling quantities, units, and measurement. Inferences in fields as diverse as the notion of beauty and fear in urban spaces are re-examined in the light of advances since the turn of millennium in both information theory and metrology, including psychometrics and entropy.

The bottom line in this broad perspective is the way artistic, linguistic, and metrological standards, like words, mediate generally navigable conceptual ideals while providing unique local creative improvisations. Faithful communication of which quantities, units, and measures are meaningful is a basic requirement irrespective of inference and challenge being tackled and irrespective of much order and complexity.

Keywords: quantity, unit, order, engineering, language, aesthetics, ordinal, entropy, metrological traceability, psychometrics, Rasch models, quality assurance

Acknowledgments: Some of the research ideas have been developed as a part of the 15HLT04 and 18HLT09 NeuroMET project consortia, which received funding (2015–22) from the EMPIR programme cofinanced by the Participating States and from the European Union's Horizon 2020 research and innovation program, and we are grateful to them and other coauthors for fruitful collaboration. Support for this work has also been provided by colleagues at RISE Research Institutes of Sweden, particularly as part of their competence platform for category-based measurement. Finally, the author is grateful to K.-D. Sommer and W. P. Fisher Jr for taking the initiative and inviting my participation in this book.

Leslie R. Pendrill, Metrology, RISE Research Institutes of Sweden

https://doi.org/10.1515/9783111036496-002

2.1 Tackling the grand challenges with more meaningful communication of quantities and units

Better interoperability, improved information exchange, more reliable trading, and ensured safety of processes and products of all kinds can be promoted by harmonized quantities, units, and their relations, thanks to their translatability from one situation to another (de Courtenay, 2015).

The need for harmonization of concepts, terminology, and language used for different types of relations among *quantity* concepts will depend on what is meaningful in each field of application when information about *quantity* is to be communicated reliably.

Harmonized quantities can be established by setting up, by consensus, a concept system, that is, a set of concepts and their relations, which can be sorted into different categories as well as form different levels and series (ISO 704:2022).

Today, where increasing amounts of information are communicated digitally and not necessarily with human intervention, consensus about a concept system can be formulated advantageously in an ontology, that is, a formal, explicit specification of a shared conceptualization, set out so that even a machine can use them for reasoning (ISO 1087:2019 clause 3.7.5; Flater, 2018):

> Example
> Data processing and electronic data interchange rely heavily on accurate, reliable, and verifiable data in databases. Both users and owners of the data have to have a common understanding of the meaning and representation of the metadata (i.e., data that describes data) (ISO/IEC 11179-31 2023). *Quantity* is one key concept with which to characterize the (meta)data.

At the same time, in support of major new legislation, for example in the new European AI Act (2023), standards, such as ISO/IEC 22989:2022, set specifications on the amount of bias (Table 2.3) and associated risks (Section 2.3.4.2) in data which need to be tested with quality-assured measurement.

A concept system for *quantity* (Section 2.1.1) will play a key role among the “top-level” ontologies, such as the recently formulated Basic Formal Ontology (ISO/IEC 21838-2:2021).

2.1.1 Relations among quantities

Concepts, such as here for *quantity,* do not exist as isolated units of information but are always in relation to each other. Two main types of relations among concepts used to establish a concept system, as will be exemplified throughout this chapter, are as follows:

- **Hierarchical** relations where different concepts are either (i) ranked conceptually on a scale from generic to specific, or (ii) related partitively in super- and subordinate relations.
- **Associative** relations, based on experience rather than hierarchy, including (iii) sequential and (iv) causal relations.

Table 2.1 has three principal columns of systematic terms, reflecting that a quantity is a **property** of an **entity** (phenomenon, body, or substance) where the property is **commensurable** (Newton, 1715, i.e., can be compared and measured, Section 2.2.3).

The various "vertical" levels in Table 2.1 reflect a hierarchy of the concept *quantity* from the generic (such as *kind of quantity*, Section 2.2.3) to the specific ("instantiations"). At each level in a hierarchy ("vertical"), different concepts may be related associatively ("horizontally").

Table 2.1: Hierarchy (generic to specific) of quantity, entity, and measurement concepts.

Level	Systematic (quantity) term (adapted from Dybkaer, 2010)	Example	Systematic (entity) term phenomena, bodies, substances	Example	Systematic (measurement) term (Pendrill, 2019)	Example
0	Kind of quantity	Space	–	–	–	–
1	Quantity, $Q = X$	Length ($x_1 - x_2$)	Kind of entity	Physical body	–	–
2	Dedicated quantity = entity quantity (*Sachgrösse*, Fleischman, 1960)	Length of rod	Entity	Rod	–	–
3	Instantiation of an entity quantity	Length of rod AB, today at 09:15	Instantiation of an entity	Rod AB, today at 09:15	Quantity as measured $Q_m = \{Q\} \cdot [Q]$	Measured length of rod AB, today at 09:15 1,234 · m

Table 2.1 is a comprehensive presentation, and in any one specific application, all levels of hierarchy and association are not necessarily included. As pointed out in Section 2.1, inclusion will depend on what is meaningful in each field of application when information about *quantity* is to be communicated reliably.

2.1.2 Human-based constructs

Measurement is conceptually distinct from *quantity* as indicated in the rightmost column of Table 2.1. Relations between them are mediated by the response of a measurement system (Section 2.3.3). Much of quality-assured measurement has developed over the decades in the context of physical science (CGPM, 2018). But there are many situations (Table 2.2) where a human being (or other agent, third column) is an active part of a measurement system instead of a technological instrument of measurement engineering. While human-based properties ("constructs") have not been included to any extent in traditional metrology, the turn of millennium has seen a substantial increase in demand for quality assurance in these areas as well the introduction of new insights and research tools to deal with categorical data, as will be illustrated throughout this chapter.

Table 2.2: Coupling item attributes to person characteristics in diverse responses for various applications of categorical data (adapted from Pendrill, 2014).

Response	**Item attribute**	**Person/probe characteristic**	**Applications (examples)**
Satisfaction	Quality of product	Customer leniency	Consumer electronics; cosmetics; health care (Rice et al., 2018); services
Performance of task	Level of difficulty of activity	(Dis-)ability	Citizen's understanding and information; learning (Melin et al., 2023); psychometrics Section 2.4.5; rehabilitation
Accessibility (e.g., of transport mode)	Barrier hinder (or cost)	Utility (or net benefit, well-being, disability, . . .)	Commuter traffic; discrete choice and valued prospects (Sundling et al., 2015; Pendrill et al., 2021, Section 2.5.3.1)
Hardness	Resistance to indentation	Ability to indent	Material characterization

Typical descriptions of constructs, for instance in the social sciences, characteristic of human-based quantities as in Table 2.2, can be found in quotes from studies of customer satisfaction, for example, in terms of "latent" variables (Melin, 2023):

> Some variables cannot be observed directly. Examples of such are intelligence, depression, suffering, attitudes, opinions, knowledge of something, satisfaction. Analysis of these variables can only be performed indirectly by employing proxy variables. The former (unobserved variables) are referred to as latent variables, whilst the latter (proxy variables) are known as observed variables. (de Battisti et al., 2010)

Note that the indirectness referred to here concerns the conceptual aspects of these quantities in themselves, rather than an indirectness – mentioned in Section 2.3.3 – which arises from the need to use a measurement system to study the constructs (Guilford, 1936).

2.2 Quantities and units of entities and their relations

2.2.1 Properties, quantities, and measures in human-based contexts

A main focus of this chapter is on human-based constructs (exemplified in Table 2.2) at certain levels of conceptual, "vertical" ranking, expressed in terms of explanatory, "horizontal" formulation (Table 2.1) of specification equations (Section 2.3.2). Before embarking on considerations of increasingly elaborate inferences (Table 2.4), from elementary syntax to the fine arts and in neuroarchitecture toward the end of the chapter, we first need to agree on what exactly measures are and how they are distinct from quantities in themselves.

Particularly where the perception and experience of various stimuli and the human response to these are of interest, the data are "categorical," that is, are results of classifications of measurement system response into categories (Section 2.3.3), even on the less-quantitative ordinal and nominal scales (Stevens, 1946). It is time these kinds of quantities – even the "latent" (Section 2.1.2) – are also included in quality-assured measurement, that is, metrology.

2.2.2 What is "measurement" and "quantity"?

Measurement is such a familiar topic that the literature contains many different and sometimes contradictory uses of the word. There are several concepts and relations that are often claimed to "define" measurement (McGrane, 2015), such as Maxwell's famous eq. (2.1) relating *quantity* to *unit*.

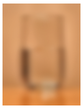
However, as shown in Table 2.3, there are many concepts and relations about measurement, such as *quantity as measured* $\mathbf{Q_m}$ which also apply equally well to quantities *in themselves,* $\boldsymbol{Q}$.

2.2.3 Commensurability of quantities and measures: Comparability and kind of quantity

Newton (1715, p. 2) wrote:

> Article I. By Number we understand, not so much a Multitude of Unities, as the abstracted Ratio of any Quantity to another Quantity of the same Kind, which we take for Unity. And this is threefold: integer, fracted, and surd: An Integer, is what is measured by Unity; a Fraction, that which a submultiple Part of Unity measures; and a Surd, to which Unity is incommensurable.

Table 2.3: Concepts associated with quantity, **Q**.

	Q		$\mathbf{Q_m}$
Concept	**Politics**	**Product**	**Measure**
Bias, error (lack of validity)	– “This vote is rigged!” – “Global warming is not Man’s fault!”	– Doesn’t start! – Clothes aren’t clean after wash!	
Precision, reliability, uncertainty	– “I didn’t say that yesterday!”	– Sometimes it doesn’t work!	
		A: Entity (product) error: $\varepsilon_p = \mathbf{Q}_1 - \mathbf{Q}_2$ for instances 1 and 2	Measurement error: $\varepsilon_m = \mathbf{Q} - \mathbf{Q_m}$

Newton’s emphasis on certain properties being “absurd” and which in the twenty-first century has again become of interest in the categorical data exemplified in Table 2.2 highlights the concept of commensurability as essential when considering the concept of *quantity*. Following Newton (1715), one could therefore define a *quantity* as the property of a phenomenon, body, or substance, which is *commensurable*.

Newton (1715) uses the word “measure” in the above citation, but in view of the ambiguity shown in Table 2.3, we have to be careful with terminology. A good definition of **measurement** is:

> Process of associating numbers, in an empirical and objective way, to the characteristics of objects and events of the real world in a way so as to relate them. (Finkelstein, 1975)

Key words in Finkelstein’s (1975) definition are: “associating numbers” and “empirical” – that is, respectively, a measurement result is a number and one needs to have made a measurement using a measurement system (in whatever form, Section 2.3.3) to qualify as a measurement (rightmost column of Table 2.3). Otherwise, quantities in themselves are the superordinate concept.

Two concepts or things are **commensurable** if they are *measurable or comparable by a common standard*. But, for instance, Sonin (1997) makes a clear distinction between comparison (Figure 2.1a) and measurement (Figure 2.1b).

Comparability in fact (without necessarily involving measurement) underlies many basic relations, such as the ability to add and subtract quantities:

- Two or more quantities cannot be added or subtracted unless they belong to the same category of mutually comparable quantities. Hence, quantities on both sides of an equal sign in a quantity equation (Section 2.2.5) shall also be of the same kind*.

To be mutually comparable, two specific properties A and B belong to the same kind, K, of quantity, that is, $A, B \in K(Q)$ if they have the same comparison and addition operations (Figure 2.1a), where:

- A comparison operation determines whether two samples A and B of the property are equal ($A = B$) or unequal ($A \neq B$)
- An addition operation defines what is meant by the sum $C = A + B$ of two samples of the property.

Some categorical data (Table 2.2), on the less-quantitative ordinal and nominal scales (Stevens, 1946), can be compared without necessarily assigning an exact number or otherwise measuring them (in Finkelstein's (1975) sense). How categorical data can be handled in measurement system analysis is presented in Section 2.3.4.2.

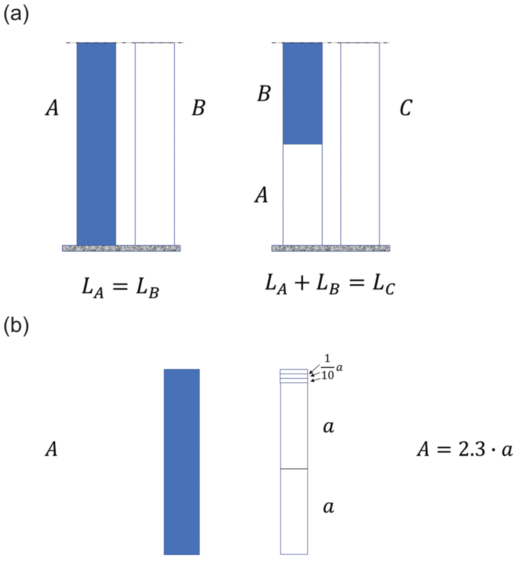


Figure 2.1: (a) Comparison of length. (b) Measurement of length (adapted from Sonin, 1997).

*Note: "Kinds of quantities are 'thought things' (German: *Gedanken-dinge*) of a different kind than quantities, they are an umbrella term (superordinate concept), namely Classes of quantities." (Fleischmann, 1960, see Table 2.1).

2.2.4 Scales of quantities and other properties: units

A quantity, Q, lies in its dimension on a scale having at least some degree of known mathematical order, such as on interval or ratio scales on which magnitudes or amounts of the quantity can be compared (Stevens, 1946).

Units are used to express how quantities scale in their respective dimensions. According to Maxwell (1871), a quantity *as measured*, Q_m, has a numerical value $\{Q\}$ expressed as the ratio:

$$\{Q\} = \frac{Q_m}{[Q]} \tag{2.1}$$

as an associative relation including a measurement unit (Q).

An interpretation of eq. (2.1) is that each scaling – in Maxwell's words: "making up" a quantity Q – involves displacing the unit $[Q]$ and counting how many times $\{Q\}$ fits into the scaled displacement, where "displacement" is not specifically in length, but in the dimension of interest. An assumption implicit in this procedure is that space is invariant in the dimension of the scaled quantity, so that the unit embodied in a scale standard does not change upon displacement. The relation with units of scale $Q = \{Q\} \cdot [Q]$ rests on the assumption of the invariance of the unit during scaling (Section 2.4.9). Where Q is instead the quantity in itself, eq. (2.1) applies equally well: in many cases, the unit of quantity will also work as the unit of measurement. (See Section 2.2.5 for further discussion of the distinction between quantities in themselves and quantities as measured.)

This interpretation of units in terms of invariance is on a more philosophical and fundamental level than a purely technical one. In physics, and several of the connections that lie behind the units of the SI system (CGPM, 2018), other more fundamental relationships between physical quantities are also used, which are expressed mathematically with equations that constitute natural laws or that define new quantities (in themselves), e.g., a physical law, force $F = m \cdot a$ (Newton's second law, if the mass, m, is constant) relating to different quantities is a universal relationship, which is applicable on all scales – from the microscopic to the cosmological.

Because of the key role of units of quantity, there must be both clear definitions of each unit and descriptions of how each unit is "realized." To be of practical use, a device not only needs to be defined but also needs to be physically realizable for the dissemination of traceability. A number of different experiments can be used to realize the definitions – called "mise en pratique." Current definitions of the units of measurement in the international system (SI) can be found in the SI brochure (CGPM, 2018a). There is no fundamental reason why the categorical data exemplified in Table 2.2 cannot be included in due course in the SI.

2.2.5 Quantities and measures

Even though a quantity refers to a commensurable property (Newton, 1715; Section 2.2.3), most quantities, Q, of interest are associated with (a) entities – phenomena, bodies, substances – *in themselves* rather than (b) those quantities Q_m *as measured* in a measurement process (Table 2.3).

An example of the advantage of making a clear distinction between quantities in themselves and as measured is as follows: In considering counts and other "quantal" quantities in quantity calculus, Flater (2023) in the context of machine readability states that "a software application that treats . . . quantities as continuous can predict outcomes that are physically impossible, such as the production of half a photon." Of course, it is physically impossible to have an object in itself that is "half a photon," but when measuring (i.e., counting) the number of photons it is quite possible to count half a photon because of limited measurement quality. As emphasized by Pendrill and Fisher (2013) in their analysis of the counting of dots by the Mundurucu Indians, the number of discrete objects to be counted can be integers trivially, but metrologically the quantities of interest are the level of difficulty in the task of counting and the ability of each counter; those two quantities are continuous attributes that together determined the probability of successful classification and are associated, respectively, with the measured object and the measurement instrument of the actual measurement system, as will be discussed further in Section 2.3.4.2.

Units have as much to do with *scaling of quantities in themselves* as to scaling of measurement responses (Section 2.2.4). Much of physics has admittedly used measurement, but that does not mean that all of physics is measurement. Stevens (1946), for instance, wrote about scaling rather than measurement units.

One can distinguish in general two types of entity-specific components of variation of the quality characteristic, Z,[1] of the object in the column for **Q** in Table 2.3 (Rossi and Crenna, 2016):

- Variable Z_{specific}: Actual variation in the quality characteristic of one specific item of the product subject to conformity assessment (e.g., changes arising when the item is used, for instance, handled in trade).
- Variable Z_{global}: Actual variation in the quality characteristic across a population of items of the product subject to conformity assessment (e.g., each manufactured item will have a different value from the other items).

Corresponding types of **measurement-specific** components of variation in the last column for $\mathbf{Q_m}$ in Table 2.3 are:

[1] Where the distinction is important, a quantity is denoted with a capital letter, for example, Z, while a lowercase letter, for example, z is used to denote the quantity value resulting from a measurement of that quantity.

– Variable Y_{specific}: Apparent variation in product, due to overall limited measurement quality when determining the value of the quality characteristic of one specific item of the product subject to conformity assessment.
– Variable Y_{global}: Apparent variation in product, due to overall limited measurement quality when determining the values of the quality characteristic of a population of items of the product subject to conformity assessment (e.g., limited sampling).

Variations associated with limited measurement quality, expressed in terms of a measurement uncertainty probability distribution function (PDF), $g_{\text{test}}(y)$ of the quantity $\xi = Y$ in the "measurement space," that is, the measurand, may partially mask observations of actual entity quality characteristic dispersion with PDF, $g_{\text{entity}}(z)$. To make clear the essential distinction between measurement variations and the quality characteristic variations that are the prime focus of conformity assessment, two different notations – Y and Z, respectively – have been deliberately chosen.

Relationships between quantities and measures are mediated by the response of a measurement system as will be discussed in Section 2.3.3.

2.3 Structural models of quality characteristics: inferences

Having identified the entity and some of its quality characteristics such as *quantity* and *measure* (Section 2.2), the next step in the quality loop is to make a **structural model** that, with various approaches, is used to formulate relations between different constructs characteristic of the quality of the entity. Any structural model can be used to summarize and validate, for both descriptive and prescriptive purposes, our knowledge and understanding of which factors determine product quality. In some cases, the structural model will enable prediction of future values of product, for instance, when planning production and designing in the widest sense.

Structural models can be formulated throughout the measurement process. This chapter is mostly about measurement, but a brief and cursory description of the production process is worthwhile, both in terms of motivating measurement as well as drawing analogies where measurement processes are regarded as a particular kind of production process (Section 2.5.4.1 and Pendrill, 2019).

Entropy and its role throughout the measurement process will be covered in Sections 2.3.6–2.3.10.

2.3.1 Inferences throughout the “measurement” process

Modeling has been described as a “fabric in the tapestry of science,” according to Rexstad (2001). In the absence of a simple “true” model, Burnham and Anderson (2002) viewed “modeling as an exercise in the approximation of the explainable information in the empirical data . . . sample[d] from some well-defined population or process . . .”

More robust inferences can be made using information-theoretic approaches compared with traditional statistical inference. Burnham and Anderson (2002) state: “Selection of a best approximating model represents the inference from the data and tells us what “effects” (represented by parameters) can be supported by the data.” Rather than continuing the traditions of statistical inference, with its “’tests’ of null hypotheses, leading to the arbitrary classification ‘significant’ versus “not significant,’” Burnham and Anderson (2002) suggest “for complex experiments . . . consideration of fitting explanatory models, hence on estimation of the size and precision of the treatment effects and on parsimony . . . a strength of evidence approach.”

We extend these information-theoretical ideas and propose to consider inference in the context of measurement system analysis (MSA) (Section 2.3.3). A measurement system is an example of a communication system. Measurement information is “transmitted” from the “source” (i.e., about an attribute associated with the entity of interest which in general also has to be formed) via a “communication channel,” that is, an instrument, to a “receiver,” that is, an operator who registers a response. The communication is completed by the operator attempting to restitute the original signal. Measurement is a “concatenation of observation and restitution” (as recalled by Bentley (2005), Sommer and Siebert (2006), and Rossi (2014)).

The MSA approach, well-established in measurement engineering, is adapted here to include situations (Table 2.2) where a human being (or other agent) acts as a measurement instrument, thereby enabling the metrology of categorical properties (Berglund et al., 2012; Pendrill, 2014b; Uher, 2018). The Rasch psychometric model constitutes restitution (of the object and instrument attributes) in such measurement-specific cases (Section 2.3.5, eq. (2.8)). This view of a measurement system is distinct from the widespread tradition in the educational and social sciences of calling a questionnaire or exam sheet an “instrument.” That point of view would admittedly be of interest if one were considering, for instance, the optimum layout to register students’ responses, perhaps when comparing pen and paper with a modern app, such as remarked by Mari et al. (2023): “a paper sheet with the printed text of a test comprising several multiple-choice items is intended to be a measuring instrument which transduces the reading comprehension ability of an individual to a response in the form of a pattern of marks on the printed checkboxes (the indication).” But it is difficult in such a remark to find what is arguably the more important cognitive task (and its level of difficulty).

As the signal is progressively transmitted from source through instrument to the operator and final restitution, the amount of information varies, as expressed in terms of entropy. Inference, based on entropy distances (Section 2.3.6), can in fact be made in order to assess how well different models – for example, of the source attribute and of the instrument – succeed at each stage in the communication (measurement) process (Figures 2.2 and 2.3).

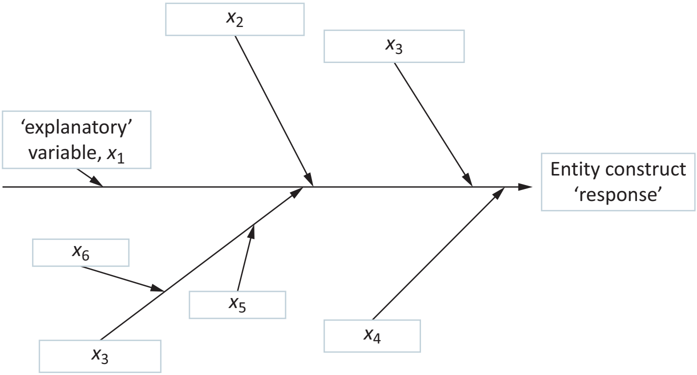


Figure 2.2: Ishikawa diagram visualizing "cause and effect" in construct specification (eq. (2.2)).

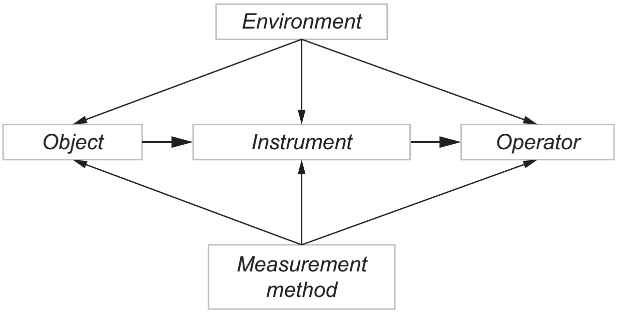


Figure 2.3: Measurement system.

In Section 2.4, such MSA multimodel inference will be developed including formation of construct specification equations (CSEs) for a number of applications, starting with the most elementary constructs. Finally, this chapter concludes in Section 2.5 with a tentative exploration of more elaborate constructs – for instance, in the fine arts, literature, and neuroarchitecture.

2.3.2 Construct specification equations and models

Modeling encompasses first setting up an expression, based on what is known about the system or entity of interest. A common visualization of explaining product in terms of "cause and effect" can be made by drawing an Ishikawa ("fishbone") diagram, exemplified in Figure 2.2, where a series of “bones” (the “causes,” one for each independent variable, x) converges to produce the overall response (“effect”), z, which depends on them.

A general CSE as a model can be formulated:

$$z = f[x_1, \ldots, x_m] \tag{2.2}$$

which involves sorting all variables into two groups – the dependent variable, z, on the LHS of eq. (2.2), and a number of independent variables, x, on the RHS, as an example of associative causal relations among quantities at any given hierarchical level in a *quantity* concept system (Section 2.2.1). A certain expression of z as a function of x (eq. (2.2)) describes how values of the entity construct z – a “response” – are related to a set of ‘explanatory’ variables x. (These expressions apply when considering the entity construct in itself. The particular case where the entity responding is a measurement system and the quantity of interest is as measured is dealt with in Section 2.3.3.)

The formulation of a CSE for an attribute of interest (**Y**, such as task difficulty or person ability, as a dependent variable, dealt with in Section 2.4.5) is often defined as a linear combination of a set, k, of explanatory (independent) variables, **X**:

$$\hat{\mathbf{y}} = \sum_k \beta_k \cdot x_k \tag{2.3}$$

As will be exemplified in Section 2.4, for memory measurements, Rasch estimates, δ_j or Θ_j, for each item, j, or person, i (in eq. (2.8)) can be the attributes of interest to be verified and validated by CSEs. The “something” that causes variation in the attribute of interest are variables that can be used to explain why some memory items are easier than others or why some persons have better memory abilities than others, i.e., the explanatory variables, $\mathbf{X_k}$.

In addition to defining the attribute of interest and identifying its explanatory variables, state-of-the-art multivariate methods for CSEs include subsequently three steps of a principal component regression (PCR) (Emardson and Jarlemark, 2005, Pendrill, 2019):

i. *Principal component analysis (PCA) among the set of explanatory variables*, $\mathbf{X_k}$: The initial set, **X**, of explanatory variables in eq. (2.3) may exhibit correlation, making it unsuitable for direct regression. PCA, where a matrix **P** of the principal components (PCs) of variation is formulated, can be used to transform **X** into an orthonormal set **X′**:

$$\mathbf{X}' = \mathbf{T} = \mathbf{X} \cdot \mathbf{P}$$

The PCs of variation are the eigenvectors, **p**, of the covariance of **X**, with eigenvalues λ:

$$\mathrm{Cov}(\mathbf{X}) \cdot \mathbf{p}_n = \lambda_n \cdot \mathbf{p}_n$$

ii. *Linear regression of the Rasch estimates, δ_j or θ_j, against* $\mathbf{X}'$ *in terms of the PCs,* $\mathbf{P}$: As a second step, the Rasch construct $\mathbf{Y}$ (eq. 2.8), e.g., task difficulty, δ, or person ability, Θ, with ε variation) is expressed as

$$\mathbf{Y} = \mathbf{T} \cdot \mathbf{C} + \varepsilon_y$$

by performing a least-squares regression against the PC:

$$\hat{\mathbf{C}} = \left(\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T}\right)^{-1} \cdot \mathbf{T}^{\mathrm{T}} \cdot \mathrm{Y}$$

iii. *Conversion back from PCs to the explanatory variables,* $\mathbf{X_k}$, in order to express the CSE for the item attribute or person characteristic is the final transforming back into the measurement space:

$$\hat{\mathbf{Y}}_0 = \mathbf{X}_0 \cdot \mathbf{P} \cdot \hat{\mathbf{C}}$$

to yield a linear combination of the explanatory variables, $\mathbf{X}$, as shown in eq. (2.3), where the coefficients in the linear predictor (CSE):

$$\beta = \mathbf{P} \cdot \hat{\mathbf{C}}$$

Thus, the formulation of CSEs includes two essential multivariate steps, equally applicable and important: First, the explanatory variables may not be the experimentally observed quantities, but some combination of these in cases where there is significant correlation between them. At step (i) in the PCR above, the procedures of multivariate analysis – such as PCA – can be used to identify the main components of variation (found by “rotating” in the explanatory-variable space from the experimental dimensions to the PC dimensions). Secondly, the CSE β-coefficients can then be determined with advantage by linear regression to the PCs (step (ii), as opposed to the experimentally observed quantities) which, together with PCA, form PCR. See further Melin and Pendrill (2023).

Design of experiments (DoEs) in traditional statistics means the process of systematically varying controllable input factors to a “manufacturing” process (in the broadest sense) so as to demonstrate the effects of these factors on the output of production (Montgomery, 1996) and is one important application where eq. (2.2) and the Ishikawa diagram (Figure 2.2) come into play, not only in manufacturing but also more broadly throughout the physical and social sciences.

A CSE can be formulated by DoE when modeling the measurement object, the entity.

Causality in memory tests when explaining task difficulty in terms of test structure (e.g., entropy, Section 2.4.5) turns out to be stronger than causal explanations of person ability in terms of biomarkers: Task difficulty depends clearly on sequence order as measured in terms of entropy, but not the other way round, that is, sequence structure is unlikely to be affected by task difficulty. For person ability, causality is less clear: a person’s cognitive ability will depend on his or her biomarkers (e.g., brain volume), but it is also conceivable that biomarker levels are to an extent determined by ability (hence the importance of longitudinal studies).

Once the causes and effects are known and "explained," then remedies for imperfections can be considered. This essentially active method of DoE can be contrasted with the more passive statistical process control (SPC). Quoting Montgomery (1996): "Statistically designed experiments are invaluable in reducing the variability in the quality characteristics and in determining the levels of controllable variables that optimize process performance."

2.3.3 Signal propagation in a measurement system

Analogous to product design, a DoE approach in measurement similar to that in statistics can be performed where one would systematically vary controllable input factors (*X*, explanatory variables) to a measurement process and register the response, *Y*. Allowance in this measurement DoE is made for both (i) variability (dispersion) – dealt with by performing analyses of variance, risk assessment and optimized uncertainty methods – and to (ii) bias (location) – dealt with by performing metrological calibration (rightmost column in Table 2.3).

Important guidance about concepts and terminology when formulating a model of the measurement process can be found in the well-established measurement engineering literature (Bentley, 2005), as follows.

A measurement system is depicted in its most elementary form in Figure 2.3.

"Measurement systems" are also mentioned but in another sense, such as in the SI system of measurement units (CGPM, 2018) and, as an example of recent terminology, Commons et al. (2014, p. 10) refer to a "measurement system" as:

> the process of associating numbers with entities or objects. . . . the components of the model (of hierarchical complexity are): – the system of entities, concatenation and comparison mathematical operators and the assignment function.

It is not, however, obvious how to relate such terminology to that used in traditional measurement engineering (Bentley, 2005; Pendrill, 2019) where, for instance, a "measurement system" is the system used for measurement: object, instrument, and operator plus environment and chosen test method shown in Figure 2.3.

Bateson (1979, p. 32) wrote: "in all thought or perception or communication about perception, there is a transformation, a coding, between the report and the thing reported, the *Ding an sich*."

An associative model (Table 2.1) of the measurement process is necessary because measurement is indirect. In the words of Guilford (1936, p. 3):

> all measurements are indirect in one sense or another. Not even simple physical measurements are direct, as the philosophically naïve individual is likely to maintain. The physical weight of an

> object is customarily determined by watching a pointer on a scale. No one could truthfully say that he “saw” the weight.

Measurement engineering provides us with a model of indirect measurements mediated by a measurement system. The output, *O*, of a measurement system (Figure 2.3), registered by an operator, is the response of the instrument to a stimulus, *I*, from the measured object, as summarized mathematically with the expression (Bentley, 2005, his eq. (2.9)):

$$O = K \cdot I + N(I) + K_M \cdot I_M \cdot I + K_i \cdot I_i + b \tag{2.4}$$

Here, sensitivity = K; nonlinearity = $N(I)$; bias = b.

Modifying disturbance = I_M, with sensitivity = K_M; interfering disturbance = I_i, with sensitivity = K_i.

Bentley (2005, his Figure 1.2) presented models of a measurement system consisting of a chain of elements, in general consisting of a mixture of up to four basic kinds: (a) sensing; (b) signal conditioning; (c) signal processing; and (d) data presentation.

Here an important fifth category of measurement system element will be added, namely (e) a decision-making element. Most measurements are not made solely for the sake of measurement, but because decisions are to be made about something (the entity) based on the measurements:

(e) **Decision-making**: algorithm producing an output on a categorical scale: the result of a decision, such as the binary, dichotomous response to, e.g., the question “is the temperature T_m below or above tolerance T_{SL}?”

$$R = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ if } \begin{bmatrix} T_m \leq T_{SL} \\ T_m > T_{SL} \end{bmatrix} \tag{2.5}$$

or a polytomous response distributed over a number of categories. Typically, decisions can be of two kinds, as in psychophysics (Iverson and Luce, 1998):

- **Identification**: T_{SL} in (2.5) is a specification limit for the quality characteristic of the entity being assessed for conformity.
- **Choice**: $T_{SL} = T_{m'}$ in (2.5) where $T_{m'}$ is a second (e.g., prior) measurement result.

2.3.4 Performance metrics of measurement systems

2.3.4.1 Traditional performance metrics

Traditional metrics with which the performance of a measurement system is rated are typically expressed in terms of measurement error: how close the system response is to the “correct” value (rightmost column of Table 2.3)?

Propagation of measurement bias and dispersion can be modeled with the following two expressions, respectively:

- Accuracy (trueness) = measured value − true value = system output − system input, $O_j - O_{j-1}$
- Accuracy (precision): $\sigma^2_{O_j} = \sigma^2_{I_{j+1}} = \left(\frac{\partial O_j}{\partial I_j}\right)^2 \cdot \sigma^2_{I_j} + \left(\frac{\partial O_j}{\partial I_{M_j}}\right)^2 \cdot \sigma^2_{I_{M_j}} + \cdots$ (2.6)

including as many terms as there are elements, *j*, of the measurement system, while assuming no correlation between the different elements. Each element of the measurement system will include a variety of measurement quantities.

2.3.4.2 Performance metrics for categorical classifications. ROC

While many measurement systems deliver responses on quantitative, continuous scales, in some cases, such as the analogue-to-digital converter and the decision-making algorithm (eq. (2.5)), the outputs will often be on discrete scales. For categorical response cases (Table 2.2), including the important decision-making response (2.5), it is not immediately obvious whether expressions such as of accuracy (eq. (2.6)) can be applied at all, since the exact mathematical distances between different categories cannot be assumed to be known (Section 2.3.5).

For these categorical responses, measurement system "accuracy" will be identified (Pendrill, 2019) with decision-making ability:

$$\text{Accuracy (decision-making)} = \text{response categorization} - \text{input (true) categorization} \tag{2.7}$$

where P_{success} is a metric of measurement system performance in terms of the probability of making the "correct" decision. In many cases, Rasch modeling (eq. (2.8)) can be done. (Classification accuracy may be limited by validity (Melin, 2023).)

For a simple binary decision (eq. (2.5)), a correct decision is described as assigning the response to the category at the output of the measurement system corresponding to the "correct" category of the measurement entity at the input to the measurement system. Analogous to the usual measurement error (eq. (2.6)), the closer the categorization, the greater the "accuracy," measured in terms of P_{success} (eq. (2.7)).

Each classification of a measurement response into a particular category is found in the approach taken in this chapter to be best treated as either identification or choice (Section 2.3.3). A related insight is that specification limits, such as dealt with in conformity assessment based on a continuous, quantitative scale, become as "marks on a ruler," thus uniting measurement of quantitative and qualitative properties. As pointed out already, the commonality between physical and social measurement (and qualitative estimations more generally) is first reached when one recognizes that the **performance metrics** of a measurement system are the same concept in both (Pendrill, 2014a,b). For this, we need to explicitly include ***decision-making*** as the third and final step – together with observation and restitution.

Account of the impact of uncertainty and decisions risks (corresponding to the pragmatic level in Table 2.4) has been included in so-called cost operating characteristics, for both testing by attribute and variable (Pendrill, 2008). In more recent work (Pendrill et al., 2023), we have re-examined and built on earlier studies of receiver operating characteristic (ROC) curves (plots of true- versus false-positive rates, (Peterson, Birdsall & Fox, 1954; Birdsall, 1974)) for the assessment of device type classification performance using an approach combining MSA and Rasch modeling (eq. (2.8)). There has been some recent work that has included modern measurement theory, demonstrating improved diagnostic performance when Rasch-compensated metrics are used to plot traditional ROC curves (Cipriani et al. (2005); Fisher & Burton (2010)). ROC curves have become an indispensable tool in machine learning, for instance. The proposed modernized ROC (Pendrill et al., 2023) goes some way to meeting the challenges posed by Linacre (1994) in his original criticism of traditional ROC curves.

2.3.5 Rasch modeling and restitution of categorical responses

The Rasch psychometric model constitutes restitution (of the object and instrument attributes) in MSA, as follows.

Using the well-established approach of measurement engineering (Section 2.3.3), the response of the measurement system to any arbitrary input stimulus value can be predicted once the various coefficients in eq. (2.4) have been evaluated by experiment in a calibration and test procedure made over a range of known input values (Rossi, 2014).

If the input signal is measurement information on a quantitative interval or ratio scale from the measurement object, then the sought-after value of the quality characteristic of the object can be deduced by a restitution process in which eq. (2.4) is inverted to estimate $I = S$ in terms of the other terms; assuming, of course, that system factors, such as the sensitivity K, remain unchanged since calibration was performed. A simple example is a measurement system where the instrument sensitivity $K \neq 1$ and there is an offset (bias), b, in the output, O. The formula for restitution of an unknown input, I, that is the stimulus value, S, of the measurement object in this case is

$$z_j = S_j = \left(\frac{R - b}{K}\right)_j = \frac{y_j - b_j}{K_j}$$

which might need to be evaluated individually at every input level if either sensitivity and/or bias vary with level.

For the categorical responses of, for instance, the decision-making elements of a measurement system (eq. (2.5)), restitution takes an analogous form, namely the Rasch psychometric model, as follows.

Categorical responses can be modeled with a measurement system where instrument sensitivity is some measure of how much a human being responds to a certain stimulus and where there can be a degree of bias when making decisions expressed as a decision-making accuracy – in terms of P_{success}, that is the probability of making a correct categorization, as in eq. (2.7).

To deal with categorical data of the kind exemplified in Table 2.2, the Rasch (1960) transformation (eq. (2.8)) is necessary in most cases, where the probability of a correct (or "successful") response (or "report") or a certain level of difficulty (or other item attribute), δ, of a task (the "thing reported") and a level of ability (or other attribute), Θ, of an instrument, is given by the Rasch (1960) expression:

$$P_{\text{success}} = \frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}} \tag{2.8}$$

Ordinality is a characteristic of categorical data – that is, a raw score, for instance, on a survey questionnaire is not in most cases a mathematically exact number, but in the best case is a monotonic trend in the right direction (higher scores mean larger quantities). Counted fractions (Tukey, 1984; Pendrill, 2019) is one main source of ordinality, characteristic of a categorical scale bound between 0% and 100%, which eq. (2.8) can compensate for. The Rasch formula also enables separate estimates of task difficulty and person ability.

Rasch in his pioneering work (1960) which has enabled metrology of human-based quantities and categorical properties in Table 2.2, wrote:

> Where this law can be applied it provides a principle of measurement on a ratio scale of both stimulus parameters and object parameters, the conceptual status of which is comparable to that of measuring mass and force. Thus, by way of an example, the reading accuracy of a child . . . can be measured with the same kind of objectivity as we may tell its weight . . .'

But one has to be careful here, not to confound physical laws (such as Newton's second law relating mass, force, and acceleration, that is, the quantities in themselves, Section 2.2.5) with the arguably more relevant measurement laws (where the latter can be found in measurement engineering, as reviewed in Section 2.3). Of course, a sensor for force can be built based on Newton's second law, but that does not mean that all sensors are based on universal physical laws. The dominating role Physics has had on the measurement field for over a century or more has admittedly colored the field. There is a tendency to be "blinded" by the elegance and universality of physics – a kind of what is popularly known these days (Nelson, 2015) as "Physics envy" if you will – and in the process missing the elementary but essential know-how of measurement systems analysis and engineering savvy (Bentley, 1995). (Rasch (1960) on another occasion correctly derives his formula in terms of the statistical Poisson distribution of the number of defects or nonconformities that occur in a unit of product when classifying it, as described in Section 2.3.5.1.)

Measurement in Physics has some unique aspects: in particular, a strong objectivity where *quantity* exists conceptually independent of any particular object and which can be estimated independently of which specific instrument is deployed (Pendrill, 2019). It is not at all obvious that such a Physics-based approach to measurement is applicable to other fields, such as the social and educational sciences. One can adopt an intermediary position by accepting a "weak" objectivity – such as the enduring appreciation of humans over the centuries for the beauty of the Mona Lisa painting (Pendrill, 2019).

In what de Courtenay (2015) has called the double interpretation of the equations of physics, various opinions have been expressed at least since the nineteenth century about the difference between *quantity* equations and *measure* equations. For instance, in contrast to the Cartesian scheme (*measure* equation), *quantity* equations "do not serve to reduce already given quantities to lines and numbers; their function is, on the contrary, to *generate* hierarchically organized quantities of different kinds" (Granger, 1988).

The distinction between *quantity* and *measure*, as discussed in Section 2.2.5, can clarify this variety of opinions. That distinction can also be useful when considering how the Rasch psychometric model relates to other approaches; regarded by some as a special case of "additive conjoint measurement" (Perline et al., 1979).

2.3.5.1 Principle of specific objectivity

Although Rasch did not use the MSA approach and terminology, the early form of his psychometric model, in response to contemporary demands for individual measures, was very much an MSA formulation. The Rasch (1960, p. xx) model (eq. (2.8)) adopted a radically different approach to statistical evaluation, as summarized in Rasch's words:

> Zubin et al. (1959) expresses: "Recourse must be had to individual statistics, treating each patient as a separate universe. Unfortunately, present day statistical methods are entirely group-centered so that there is a real need for developing individual-centered statistics."
>
> Individual-centered statistical techniques require models in which each individual is characterised separately and from which, given adequate data, the individual parameters can be estimated . . . . Symmetrically, it ought to be possible to compare stimuli belonging to the same class – "measuring the same thing" – independent of which particular individuals within a class considered were instrumental for the comparison.

This separation of stimuli and individual's attributes is key in metrology and is particularly important to consider when assuring metrological quality of performance tests and other classification data (Pendrill, 2014b). The separation has many similarities to the basic idea in engineering metrology, where calibrating and correcting for known errors in an instrument separately from whatever object is being measured is essential for the establishment of measurement standards for traceability in quality-assured measurement.

A recent use of the Rasch approach is by Commons et al. (2014, p. 9) in their model of hierarchical complexity (Section 2.4.3):

> Previous theories of stage have confounded the stimulus and response in assessing stage by simply scoring responses and ignoring the task or stimulus. The model of hierarchical complexity separates the task or stimulus from the performance.

Commons' et al. (2014, p. 10) explain further:

> the entities are task actions of organisms, social groups, and computers (Krantz, Luce, Suppes, & Tversky, 1971) . . . (and) actions are defined as behavioral events that produce outcomes. . . . A task can be defined as a set of required actions that obtain an objective, though the performed actions may or may not complete a given task. The study of tasks appears in psychophysics, a branch of stimulus control theory in psychology (Green & Swets, 1966; Luce, 1959) and in artificial intelligence. (Goel & Chandrasekaran, 1992)

Rasch's (1961) model posits that the odds ratio of successfully performing a task is equal to the ratio of an ability,[2] h, to a difficulty, k:

$$\frac{P_{\text{success}}}{1-P_{\text{success}}}=\frac{h}{k}$$

where the test person ("agent") ability, $\theta=\log(h)$, and task ("object") difficulty, $\delta=\log(k)$ (in eq. (2.8), or other item:probe pairs of attributes, Table 2.2). The quality of each response is rated in a similar way to any product, in terms of the number of defects or nonconformities or "unsuccessful" responses. The original Rasch (1961) formulation referred to a probabilistic Poisson distribution, well known from quality control as a model of the number of defects or nonconformities that occur in a unit of product when classifying it:

$$p(x)=\frac{e^{-\lambda}\cdot\lambda^{x}}{x!};\; x=0,1,\,\ldots$$

where x is the quality characteristic being classified. The parameter λ is equal directly both to the mean and variance of the Poisson distribution. In Rasch's (1961) model,

$$\lambda=h^{-1}\cdot k$$

In accordance with Rasch's principle of specific objectivity, task difficulty and agent ability need to be treated separately (although the response P_{success} depends on both).

[2] Rasch (1961) used the person attribute "inability" instead, given by h^{-1}.

2.3.6 Entropy, quantities and the measurement process

In the rest of this section, and in many of the applications recounted in Sections 2.4 and 2.5, entropy turns out to be a dominant explanatory variable.

Both individual entity constructs as well as the passage of information when regarding measurement as a particular kind of information communication (Section 2.3.6.3) can be explained with entropy. This includes faithful description of the observation process and making reliable and valid inferences about different models (Section 2.3.8).

2.3.6.1 Increasingly "meaningful" messages

The amount of information transmitted from the measurement object to the observer can range from a simple signal through to increasingly "meaningful" messages, as is captured in four levels of increasing richness in information theory (Weaver and Shannon, 1963; Klir and Folger, 1988) as given in the first column of Table 2.4. Depending on what kind of meaning is to be communicated, the kind of (measurement) information will fall into one or other of the extended *quantity* calculus hierarchy (Section 2.2.1, second column of Table 2.4).

Table 2.4: Comparing concepts in information theory and quantity calculus (Pendrill, 2019).

Information theory	Quantity calculus
Effectiveness – "changing conduct": relationship between signs of communication and actively "improving" the entities they stand for (Weinberger, 2003)	Nature of quantity (Emerson, 2008)
Pragmatic – "utility": relationship between signs of communication and their utility (value, impact)	Kind of quantity (Fleischmann, 1960)
Semantic – "meaning": relationship between signs of communication and entities they stand for	Quantity
Syntax – "signs": relationship among signs of communication	Value of quantity

Weaver (1949) illustrated communication of information at the lowest (syntax, Table 2.4) level with the following vignette:

> An engineering communication theory is just like a very proper and discreet girl accepting your telegram. She pays no attention to the meaning, whether it be sad, or joyous, or embarrassing.

2.3.6.2 Amount of “useful” information: entropy

As has become popular in information theory, the concept of entropy is of interest as a measure of the amount of information. To consider entropy in depth, it is productive to trace the roots of the concept. Before the term “entropy” appeared in thermodynamics, Carnot (the elder) (1803) stated that:

> In any machine the accelerations and shocks of the moving parts represent losses of moment of activity . . . . In other words, in any natural process there exists an inherent tendency towards the dissipation of useful energy.

Ideas about “useful” information, analogous to useful energy in the rattling machines of the early Industrial Revolution, can be developed to model how well any task of a certain difficulty is performed by an agent of a certain ability in terms of entropy.

Early measures of information referred to different aspects of ambiguity which can make tasks more difficult or persons less able:

- Hartley (1928) information: $I(N) = \log_s(N)$
  based on classical set theory, pertains to non-specificity, where N is the total number of alternatives, that is, $I(N)$ is the amount of information needed to characterize one of N alternatives (Klir and Folger, 1988).
- Shannon (1948) entropy:

$$H(p) = -p \cdot \log_s(p) \tag{2.9}$$

  formulated in terms of probability (p) theory, to conflict or dissonance in evidence (Klir and Folger, 1988).

There is an extensive literature in which the validity of analogies between the entropy concept in information theory and thermodynamics is considered in depth, see for example, Maroney (2009). A summary about making these analogies is given in Wiki pedia (2021), where statements can be found, such as: “Ultimately, the criticism of the link between thermodynamic entropy and information entropy is a matter of terminology, rather than substance. Neither side in the controversy will disagree on the solution to a particular thermodynamic or information-theoretic problem.”

In Section 2.4 we give an account of the use of entropy as an explanatory variable in neuropsychological assessments of cognition. With the understanding gained from formulating CSE for task difficulty, better metrics of cognition can be formed by carefully combining selected items from legacy short-term memory tests so as to enhance coherence in item design while not jeopardizing validity (Melin et al., 2022a, 2023a). It enables metrological references in cognitive memory tests analogous to certified reference procedures for metrology in chemistry (Pendrill, 2019) and Section 2.4.7.

The concept of entropy has wide applicability as will be explored in the final section. For example, it can also explain the efficiency of any organization in terms of entropy-based measures of synergy (the case of hospitals: Chen et al., 2015, Pendrill, Espinoza et al., 2021) or how well human muscles can deliver force ergonomically, for instance when climbing stairs or lifting heavy packages (Maršik & Mejsnar, 1994). A CSE for task difficulty has some similarities to creating artwork algorithmically (Nake (1974)). Section 2.5 will also explore the perception of built environments.

2.3.6.3 Entropy throughout the measurement process

The amount of "useful information" in a measurement system, by analogy to a certain extent with the original entropy concept as a measure of "useful energy" in steam engines (Section 2.3.6.2), can be described with the well-known conditional entropy expression:

$$H(Y|Z) = H(Z,Y) - H(Z) \tag{2.10}$$

Expression (2.10) describes how the amount of information (depicted in Figure 2.4) changes during transmission in a measurement system (Figure 2.3), in terms of the entropy in the response (*Y*) of the system when observing a quantity (*Z*) attributed to the measurement object.

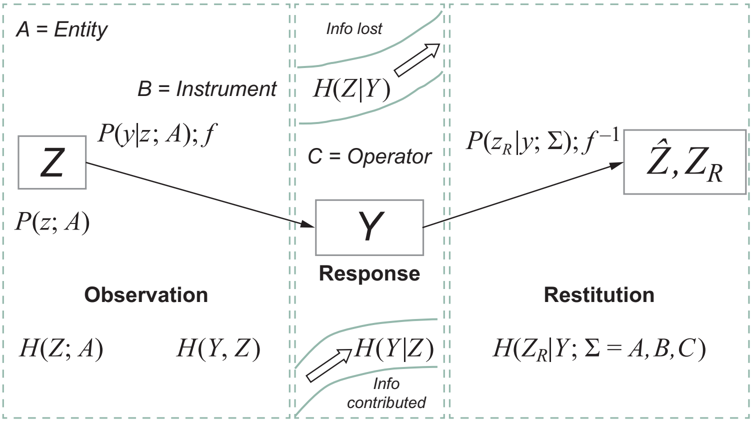


Figure 2.4: Entropy lost and gained in a communication (measurement) system.

Each attribute of the various elements (object, *A*, instrument, *B*, operator, *C,* etc.) of a measurement system can be explained causally in terms of construct specification equations (CSE, Section 2.3.2). This is done based on our best understanding of each construct. Entropy turns out to be a dominant explanatory variable throughout all stages of the measurement process (and as will be seen in the various applications recounted in the remaining sections of this chapter).

At the start of the measurement process, there is an initial “deficit” in entropy (i.e., “surplus” information) $H(Z)$, coming from prior knowledge (prior distribution P) of the measurand (attribute, Z, of object, entity A). Losses and distortions $H(Z, Y)$ increase the entropy through measurement imperfections (including measurement system attributes, such as the sensitivity and resolution of an instrument (B)), leading finally to a posterior distribution (Q) in the response Y with entropy $H(Y|Z)$ as the result of the measurement process as registered by an observer (C). The notation here is analogous to that used by Rossi (2014) in a probabilistic model of the measurement process, and as described in more detail in Melin et al. (2022).

2.3.7 Differences in entity, response, and measured values: Entropy and histogram distances

Distances on scales of categorical data are not immediately quantitative (Section 2.3.5). Research in among others modern image processing has shown the possibilities of making measures of histogram distances in terms of entropy, as follows.

The amount of measurement information on the categorical scales of signals at any one point and state in the measurement process is in general the summed (change in) entropy, which for a discrete PMF is $\Delta\mathbb{H}(Q) = -\sum_c q_c \cdot \ln(q_c)$, where q_c is the occupancy of category c. Information in each classification category, c, is expressed as a Shannon surprisal $-\ln(q_c)$ (eq. 2.9), while the relative contribution to the total entropy is weighted with the relative occupancy, q_c (Pendrill, 2019, chapter 5); see Figure 2.5.

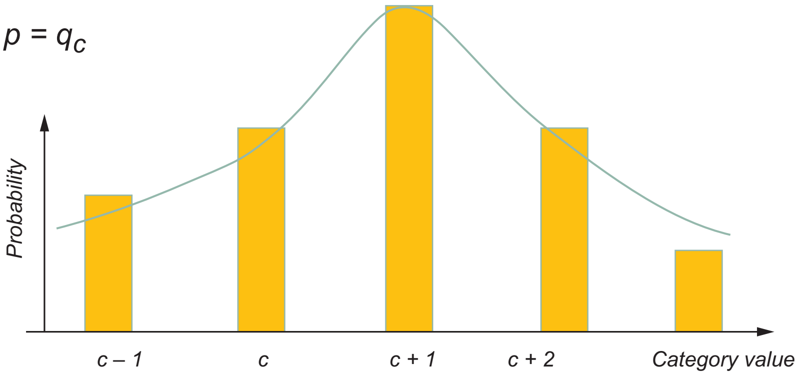


Figure 2.5: Entropy (amount of information): $\Delta\mathbb{H}(\boldsymbol{Q}) = -\sum_c \boldsymbol{q}_c \cdot \ln(\boldsymbol{q}_c)$ on a categorical scale.

Pele and Werman (2010), among others, investigated several different histogram distance metrics, including an entropy-based measure of interhistogram distances on a *semantic* scale, which is a variant of the Kullback-Leibler (KL) (1951) distance:

$$d_{KL}(Z,Y) = \sum_c z_c \cdot \ln_b\left(\frac{z_c}{y_c}\right)$$

for a response Y to an object attribute Z.

As is well known, the KL distance is not considered a complete metric since in general it is not symmetric, that is, $d_{KL}(Z,Y) \neq d_{KL}(Y,Z)$. Because the KL distance is not strictly a metric, alternative measures need to be sought where the variant of d_{KL} after Jeffrey is (Pele and Werman, 2010 and references therein):

$$d_J(Z,Y) = \sum_c \left[z_c \cdot ln\left(\frac{z_c}{m_c}\right) + y_c \cdot \ln\left(\frac{y_c}{m_c}\right)\right]$$

where $m_c = \frac{z_c + y_c}{2}$.

In its infinitesimal form the Kullback-Leibler (1951) distance is however a metric tensor: the Fisher information metric, $g_{j,k}(\theta_0)$, which appears to second order in a Taylor expansion of the KL distance, on small displacements $\Delta\theta^j = (\theta - \theta_0)^j$:

$$d_{KL}(P(\theta)|P(\theta_0)) = \Delta\theta^j \cdot \Delta\theta^k \cdot g_{j,k}(\theta_0) + \cdots$$

The Fisher information metric, as a Hessian matrix of the divergence,

$$g_{j,k}(\theta_0) = \frac{\partial^2}{\partial\theta^j \cdot \partial\theta^k}_{\theta=\theta_0} d_{KL}(P(\theta)|P(\theta_0))$$

then enters for example in the Wald test statistic $z = \frac{\hat{\theta} - \theta_0}{\mathrm{SE}}$, where $\mathrm{SE} = \frac{1}{\sqrt{\frac{\partial^2}{\partial\theta^j \cdot \partial\theta^k}_{\theta=\theta_0} d_{KL}(P(\theta)|P(\theta_0))}}$. In terms of entropy $(z,\theta) = -\ln[p(z,\theta)]$, $g_{j,k}(\theta_0) = \int \frac{\partial^2}{\partial\theta^j \cdot \partial\theta^k}_{\theta=\theta_0} H(z,\theta) \cdot p(z,\theta) \cdot dz = \mathbb{E}\left[\frac{\partial^2}{\partial\theta^j \cdot \partial\theta^k}_{\theta=\theta_0} H(z,\theta)\right]$. An equivalence can be set between the subjective distance $D_{KL}(a,b)$ and the conditional entropy $H(Q|P)$.

2.3.8 Entropy, perception and decision-making

Considering the decision-making process as part of the transmission of information in a measurement system as a perception of pairwise discrimination of adjacent stimuli, such as in choice in cognitive psychology (Iverson & Luce, 1998, eq. (2.5)), the subjective (Kullback-Leibler (KL) (1951)) distance $D_{KL}(a,b)$ between two stimuli, a and $b > a$, is expressed (Dzhafarov, 2012) as the integral over the level, s, of stimulus of a measure of the ability to perceive a dissimilarity:

$$P(s, s+ds) = Pr[b \,\text{is judged to be greater than}\, a]:$$

$$D_{KL}(a,b) = \int_a^b \frac{P(s,s+ds)}{ds} \cdot ds$$

The subjective distance, *D(a,b)*, reduces to Fechner's law used widely in psychophysics:

$$D(a,b) = k \cdot \log\left(\frac{b}{a}\right)$$

when the gradient of the dissimilarity $D(s,s+ds)/ds = (k/s)$ and a is set to the "absolute threshold."

This approach of relating subjective distance to accumulated dissimilarity in terms of discrimination probabilities can be extended to include not only continua of the senses (of colors, sounds, etc.) but also to Fechnerian scaling of the perception of discrete object sets where stimulus sets are "isolated entities," such as schematic faces; letters of the alphabet; dialects and the like (Nerbonne et al., 1999).

In the simplest, dichotomous case where the prior is known, such as in the elementary case of counting dots (Pendrill & Fisher, 2013), the subjective distance $D_{KL}(a,b)$ between the distributions (*P*) and (*Q*) for the two stimuli *a* and *b* to the measurement-based decision is obtained by substituting $P(s,s+ds)/ds = dP_{\text{success}}$ and $ds = -z$ to yield:

$$\begin{aligned} D_{KL}(P,Q) = \int -z \cdot dP_{\text{success}} = \\ -\left[P_{\text{success}} \cdot \log(P_{\text{success}}) + (1-P_{\text{success}}) \cdot \log(1-P_{\text{success}})\right] \\ = H(P,Q) - H(P) = H(Q|P) \end{aligned} \tag{2.11}$$

where we set an equivalence between the subjective distance $D_{KL}(a,b)$ and the conditional entropy $H(Q|P)$. How measurement information is acquired, transmitted, lost, and distorted on transmission through the measurement system and in communication more generally can be described in terms of entropy (both for measurement units and uncertainty) as discussed in Section 2.3.6.

A straightforward derivation of decision probabilities can be made with the method of Lagrange multipliers subject to the constraint of maximizing entropy, leading readily (Linacre, 2006; Pendrill, 2019; Massof et al., 2023) to the logistic regression link function:

$$z = \log\left[\frac{P_{\text{success}}}{1-P_{\text{success}}}\right]$$

A particular form of this logistic equation is the Rasch measurement model (Section 2.3.5), where $z = \theta - \delta = \log(P_{\text{success}}/(1-P_{\text{success}}))$ that can be applied to transform ("restitute") the ordinal, "counted fraction" data P_{success}, onto the more quantitative scale for *Θ* and *δ*. In our MSA approach (Section 2.3.3), the object – such as a task posed in a test item – is distinguished from the instrument – here, the agent tackling the task and responding “yes” or “no.” The Rasch (1960) psychometric approach (eq. (2.8)) enables – according to his principle of specific objectivity – this distinction.

2.3.9 Information criteria in making inferences on models

Model selection based on information theory represents a quite different approach in the statistical sciences, and the resulting selected model may differ substantially from model selection based on some form of statistical null hypothesis testing (Burnham & Anderson, 2002).

Various model-selection criteria, including Akaike’s information criterion (AIC), corrected AIC (AICc), and the Bayesian information criterion (BIC), can be useful in providing a balance between “reproducing the data and avoiding overfitting” [Murari et al. (2019)]. As recalled by Murari et al. (2019): “The BIC criterion is derived in the framework of Bayesian theory and it is meant to maximize the posterior probability of a model given the data. AIC is based on the Kullback-Leibler Divergence (section 2.3.7) and essentially estimates the information lost by a given model. Therefore, it is assumed that the less information a model loses, the higher its quality”

$$\mathrm{BI}C = -2\cdot\ln(\mathbb{L}) + k\cdot\ln(n) \tag{2.12}$$

$$\mathrm{AIC} = -2\cdot\ln(\mathbb{L}) + 2\cdot k$$

where $\mathbb{L}$ is the likelihood of the model given the data, k the number of estimated parameters in the model, and n the number of entries in the database.

Because the likelihood of a model is not easy to calculate, typically one assumes that the “model and data errors are identically distributed and independently sampled from a Normal distribution” (Murari et al., 2019) in which case:

$$\mathrm{BIC} = n\cdot\ln\left(\sigma_\varepsilon^2\right) + k\cdot\ln(n),$$

where σ_ε^2 is the variance of the residuals.

Similarly,

$$\mathrm{AIC} = n\cdot\ln(\mathrm{MSE}) + 2\cdot k,$$

where MSE is the mean-squared error of the residuals.

The root-mean-square residual (RMSR) for each n being formed, where

$$\mathrm{RMS}R(n) = \left[\frac{\chi^2_{\mathrm{obs}}(n)}{m-n-1}\right]^{1/2}$$

which applies for $n < m-1$. Here $\chi^2_{\mathrm{obs}}(n)$ is the sum of squared residuals.

$$\text{AIC\_ln} = m \cdot \ln(\text{RMSR}(n)) + 2(n+1)$$

$$\text{BIC\_ln} = m \cdot \ln(\text{RMSR}(n)) + (n+1) \cdot \ln(m)$$

Adding to Murari et al.'s (2019) statement that:

> the use of the Shannon entropy in selection criteria . . .based on the observation that, if a model were perfect, the residuals should be due only to the noise affecting the data,

we can, with our MSE interpretation (Section 2.3.3), consider inferences about how well one succeeds in modeling the response (*Q*) to information from the original source (P) in eq. (2.11) *at every stage* in the measurement process, as introduced in Section 2.3.1. In any measurement process (as in a general communication system), information can be both lost or gained, as illustrated in Figure 2.4. There may be both distortion (bias), uncertainty (less clarity) through dissipation of useful information as well as additional information gained in the process, for instance from prior knowledge and background information. The distinction between measurement errors and quantity errors also needs to be made (Table 2.3).

2.3.10 Uncertainty

Informational entropy, as a measure of the amount of information, is a key concept when seeking metrological quality assurance of measurement systems, both in terms of measurement uncertainty (loss of information) and of traceability (distortion of information, Table 2.1). Two distinct contexts can be identified, where one seeks stationary (maximum or minimum, respectively) values of the entropy:

- Although entropy can be both lost and gained in the process, the net change in overall entropy for the *whole* system on transmission of measurement information cannot decrease, thus allowing realistic estimates of measurement uncertainty, in line with the second law of thermodynamics;
- the best units for metrological traceability are those with the most order, that is, least entropy, as an example of the principle of least action.

In the limit where discrete multinomial expressions invoked for the entropy of categorical responses go toward the familiar continuous scales of traditional uncertainty presentations (Pendrill, 2019, chapter 4):

$$\Delta\mathbb{H}(Q) = -\int_{-\infty}^{\infty} p(Q) \cdot \ln(Q) \cdot dQ = \ln\left[\sqrt{2\pi} \cdot u(Q)\right] + \frac{1}{2} \tag{2.13}$$

giving the integrated probabilistic (Shannon, eq. 2.9) formulation of the entropy across the width ("uncertainty") in the response. Equation (2.13) indicates that two separate approaches – (i) standard uncertainty, *u*, and (ii) decisions risks – can be unified.

The relation $\Delta\mathbb{H}(Q) = \ln\left[\sqrt{2\pi}\cdot u(Q)\right] + \frac{1}{2}$ applies to the particular case where the probability distribution function of the outcome Q is taken to be Gaussian (normal), that is, $p(Q) = N\left[\bar{Q}, u(Q)\right]$. Inversion of eq. (2.13) suggests an alternative expression of measurement uncertainty:

$$u_q \sim e^{\Delta H(Q)} \tag{2.14}$$

more akin to the concepts of information theory than the classic standard uncertainties [JCGM GUM].

We have a certain preference to express uncertainty in terms of an increase *ΔH* in entropy instead of a standard deviation (Zidek & van Eeden, 2003) because it is conceptually closer to "uncertainty" in everyday language – "decision quandary"; is also substantially distribution-free; and is indeed accessible to treatment not only with probability theory but also with possibility and plausibility theories.

2.4 Models of language and tasks: relative proportions of order and complexity

In this section, quantities, units, and related concepts presented so far in this chapter – such as entropy, symmetry, conservation – will be applied in various ways to effective and meaningful communication of information (mostly syntax but also some "words") by measurement.

2.4.1 Entropy as an explanatory variable

Entropy will be found to be a measure of "useful" information (Section 2.3.6) in many applications wherever an amount of information is transmitted from the measured object to the observer, ranging from a simple signal to increasingly "meaningful" messages, as captured in four levels of increasing richness in information theory, from syntax upward (Table 2.3, Weaver & Shannon, 1963; Klir & Folger, 1988).

The challenge in most cases will be to formulate what represents a bit of information rather than how many bits there are (Section 2.4.3).

2.4.2 "Words," order, complexity, and entropy

Considering the communication of measurement information at any level, a useful starting point for introducing the concept of "units" (Section 2.2.4) is to recognize the analogical role of words for effective linguistic communication. Metrological traceability provides the necessary measurement comparability, and calibration means being able to trace measurement results to measurement standards that, so to speak, embody defined “recognizable” or “meaningful” quantities of the unit.

A classic example illustrates different information content in the following three messages consisting of an equal number of symbols (digits or letters):

$$\begin{gathered}\text{“100110001100”}\\ \text{“agurjerhjjkl”}\\ \text{“this message”}\end{gathered} \tag{2.15}$$

It is obvious to anyone who understands English that the third message conveys more information than the other two messages, although all three messages have the same number of characters, reflecting how efficiently each string can communicate meaning.

In the context of information technology, a concept of “objective complexity” has been associated for some time with a string of binary digits (e.g., “0” or “1”), as the length of the shortest computer program required to generate the string, as formulated in the early 1960s by Kolmogorov (1965) and Chaitin (1969). The adjective “objective” in this context meant for those authors that the complexity could be calculated algorithmically. The Kolmogorov complexity has also been called *algorithmic information* and *algorithmic randomness*. A recent example of using the Kolmogorov compression complexity was in differentiating icons in the visual arts (Peptenatu et al., 2022).

As reviewed in Section 2.5.4, work in the field of digital image processing and recent advances in finding automated measures of visual complexity has, according to Marin and Leder (2013), profited the study of subjective complexity and its relations to aesthetic experience.

2.4.3 “Horizontal” complexity

As mentioned in the previous section, complexity is one concept that has been invoked in several studies with wide applicability but where its meaning has been variable. Commons et al. (2014), for instance, in their model of hierarchical complexity, claim that “complexity” alone forms the basis for ranking actions and tasks in educational contexts. They refer to the (horizontal) complexity of an action, that is, the sum of n bits required by tasks that require “yes-no” questions where the number of actions is 2^n. The amount of information in Commons et al. (2014) formulation of $N = 2^n$actions is readily calculated with the Hartley (1928) expression (Section 2.3.6.2) as: $I(N) = \log_2(2^n) = n$ bits. Naturally “task sequences of task behaviors form hierarchies that become increasingly complex . . . (where) less complex tasks must be completed and practiced before more complex tasks can be acquired” (Commons et al. 2014, p. 9). Overall development is obviously more complicated (Dawson-Tunik, et al., 2005).

Further examples may be found in the study of symmetry in molecular or crystalline formations (Schneider et al., 1986, 1990). Schneider et al. (1990) describe Shannon's measure of information (eq. (2.9)) as indicating how much choice is involved in a particular selection from among two or more alternatives. Measurements are given in bits, where one bit is the amount of information necessary to select one of two possible states (e.g., a yes-no question) in a binary choice or more generally $\log_2 M$ bits required to select one possibility from M possible states. Shannon's amount of information (eq. (2.8)) is related to the *decrease* in the number of possibilities. Schneider et al. (1990) give an example of four possibilities, and one only answers one yes-no question, then two possibilities remain, so the information gained in $\log_2 4 - \log_2 2 = 1$ bit.

As a word of warning, Andersson and Törnberg (2018) when tackling "irreducibly complex" societal systems, emphasize that complexity is one only of a number of more or less closely related concepts: "'Complexity' usually does not point at any particular idea about complexity, nor at any particular generating process, but works mostly as a catch-all term for problems that overwhelm us in some sense; things like massive parallelism, multilevel hierarchization, heterogeneity, tangled 'seamless webs', emergence, nonlinearity and sensitivity to disturbances, or combinations thereof." They argue that the pair of concepts – complexity and complicatedness – can be plotted as two orthogonal axes, where minimum values of both refer to simplicity while maximum values correspond to wicked societal systems.

2.4.4 Order, entropy, and combinatorics

As is evident from the different "readability" and "information content" of the messages (2.15), the challenge is to formulate what represents a bit of information rather than how many bits there are. Analogies can be drawn with decomposing a message into words (and other structures – syntax, semantic and pragmatic) and a measurement result in a set of units, as recounted below. The concept of "chunking" was introduced early by Miller (1956).

Brillouin (1962) emphasized is that it is not simply the number of symbols, but rather the number of *combinations* of symbols that determines the amount of "useful" information (Section 2.3.6.2) in a message. According to Brillouin (1962), a "message" to be communicated can be characterized as an amount of information, I, which in general depends on the number, N_j, ($j = 1, \ldots, M$) of symbols of M different types distributed over a number, G, of categories (or cells) $G = \sum_{j=1}^{M} N_j$. In the simplest case (with no repeats, i.e., $N_j = 1$, for all symbols) combinatorics dictates that:

$$I \sim \ln P \sim K \cdot \ln(G!) \tag{2.16}$$

where $K = \frac{1}{\ln(b)}$ is a normalization constant and b is the base of the logarithm according to Brillouin (1962, chapter 1, section 2: *unit systems*).

Brillouin starts with the simplest case, considering the number of ways of filling each of G cells with either 0 or 1, but never both. He refers to this as the same as a problem in Fermi statistics in physics, where two elementary particles (of the kind caller "fermions," such as electrons) cannot be present in a given quantum state (a "cell" or category) at the same time. The number of ways, P, one can fill the G cells is equal to the number of ways N_0 cells can be filled with 0, since once N_0 0's have been distributed, then the remaining N_1 cells must each contain 1:

$$P = \frac{G!}{N_0! \cdot N_1!} \tag{2.17}$$

This is the number of "messages" of G symbols, consisting of one symbol of a two-letter alphabet used N_0 times and the other symbol used N_1 times. The amount of information in one of these messages is using the Shannon entropy formula (eq. (2.9)):

$$H(Z) = K \cdot \ln P = K \cdot [\ln(G!) - \ln(N_0!) - \ln(N_1!)]$$

Generalizing to an n-letter alphabet, the probability of encountering the jth symbol is $p_j = \frac{N_j}{G}$, which can be summed to unity. The total number, P, of messages that can be obtained by distributing the symbols at random over the G cells (with never more than one symbol per cell) is $P = \frac{G!}{\prod_{j=1}^{M} N_j!}$. The information theoretical "Shannon" entropy (eq. (2.9)), which is a measure of the amount of information in these messages, is then given by the classic Brillouin expression (1962):

$$H(Z) = K \cdot \ln\, P = K \cdot \left[\ln(G!) - \sum_{j=1}^{M} \ln\left(N_j!\right) \right] \tag{2.18}$$

where K is a normalization constant (defined in eq. (2.16)).

Note the differences in how expression (2.18) contains the combinatoric factorial terms, such as $G!$ and $N_j!$ compared with Sigaki et al. (2018) who attempted to explain the degree of visual order of artworks in terms of entropy, where the number of permutations in the latter occurs instead as the number, $n = \left(d_x \cdot d_y\right)!$, of terms summed in the Shannon entropy expression (eq. (2.9)).

2.4.5 Entropy and CSE in memory tests: task difficulty in immediate serial recalling

Entropy has a broad applicability when formulating CSE in general (Melin & Pendrill, 2023) and is simply and generally described as a measure of the amount of "useful" information or "useful" energy (Section 2.3.6.2): higher entropy implies less order leading to a loss of information or energy, and vice versa, lower entropy implies higher order and less uncertainty.

2.4.5.1 Block, digit and word recall tasks and entropy

Brillouin's (1962) eq. (2.16) is the basis for our approach to modeling task difficulty in terms of the number of different combinations of symbols to be tackled in the task by a human instrument in a series of studies of immediate recall for a number of elementary memory tests where sequences consist of blocks, digits, words (2.4.5.2 and 2.4.5.3) at the lowest (syntax) conceptual levels of the information hierarchy (Table 2.4) (Pendrill, 2019; Melin et al., 2023). The "task" in our MSA approach is the measurement object (where typically task difficulty is the attribute, i.e., the quantity *in itself* characteristic of the object). The difficulty of a task is posited to be proportional to its entropy – or more-ordered task will be easier. Thanks to the principle of specific objectivity, application of the Rasch model (eq. (2.8)) enables task difficulty to be estimated separately from person ability (Section 2.3.5.1).

This entropy-based approach to explaining recall task difficulty has led to significant improvements in the efficiency and reliability of a number of legacy memory tests (such as the Knox (1914) cube test (KCT), Corsi block test (CBT), digit span test (DST), and word list tests (RAVLT)). It is remarkable that basically the same formula for task difficulty $\delta_j(\text{logits}) = \beta \cdot \text{Entropy}_j$ applies to all tests, where the entropy of each item, *j*, is calculated with Brillouin's (1962) eq. (2.16), simply inputting the number, *G*, of objects recalled as the characteristic of the sequence structure: Pearson correlation coefficients of at least $R = 0.9$ were achieved when explaining task difficulty in the block tests (KCT and CBT), $\beta = +1.2(6)$ and the digit test (DST), $\beta = +1.0(2)$ plus an offset of $-6(3)$ logits, where the numbers in brackets are the ($k = 2$ coverage factor) expanded measurement uncertainties.

The new European NeuroMET memory metric (NMM) is composed of items selected from these legacy tests where the selection has been guided by the entropy-based theory of task difficulty presented here and has been demonstrated to be more efficient and with much reduced measurement uncertainties compared with the individual legacy memory tests (Melin et al., 2023).

2.4.5.2 Learning and entropy

Even the effects of learning could be simply modeled: on each repeat of the 15 words to be recalled in 5 repeated trials of the word learning list test *RAVLT IR*, the task difficulty $\delta_{Mr}(\text{trial}) = \frac{\delta_{Mr,0}}{\sqrt{\text{trial}}}$, where $\delta_{Mr,0}$ is the difficulty of the first trial, in much the same way as the standard deviation of a series mean reduces with the inverse root of the number of repeats (i.e., number of degrees of freedom) (Melin, Kettunen et al., 2022). Similar learning effects had been found in Cordier et al.'s (1994) study of a complex motor behavior – specifically, a constrained free climbing task – where the main concept was again found to be entropy as a measure the degree of structuring of a complex task. They found that the entropy of the trajectory decreased as learning progresses, and that the shape of the entropy curve is a function of the climber's level of expertise.

2.4.5.3 Serial position effects: primacy and recency

The word learning list test *RAVLT IR* provides a conceptually simple example with which to test our theories further – being one step more advanced semantically than our previous syntax studies of nonverbal, block, and number recalls. Over and above the inherent difficulty of recalling any symbol in a sequence (Section 2.4.5.1), a peculiarity of *RAVLT* is serial position effects (SPE): that it is expected that, all other factors being equal, the initial and final symbols in the word list should be somewhat easier to recall than the symbols in the middle of the sequence according to the well-known effects of primacy and recency. SPEs are evident in several memory recall studies, as noted in the review of Hurlstone et al. (2014):

> Forward accuracy serial position curves exhibiting effects of primacy and recency are not confined to verbal memoranda. The forward serial position curves associated with the recall of sequences composed of various types of nonverbal stimuli have been shown to exhibit an extensive primacy effect accompanied by a one-item recency effect. These stimuli include visual–spatial locations, visual–spatial movements, auditory–spatial locations, visual matrix patterns, and unfamiliar faces.

The same basic entropy model as in Brillouin's (1962) eq. (2.18), which accounts for the reduction in task difficulty when symbols are repeated (N_j times), can also be used to explain SPE for the verbal tests (L words) *RAVLT IR*; that is, the fact that it is easier to recall words from the beginning and the end of a list, j:

$$\delta_{j,Pr} = -M \cdot \ln\left(G_j!\right);\ G = \text{ item position}$$

$$\delta_{j,Rr} = -M \cdot \ln\left(G_j!\right);\ G = L - 1 - \text{ item position}$$

where $M = \frac{1}{\ln(L)} = \frac{1}{\ln(15)} = 0,369$ for a sequence of length, $L = 15$ words.

The basic task difficulty of recalling any word in the list is reduced for each SPE by the corresponding reduction in entropy according to this pair of equations for *primacy* and *recency*, respectively.

As in the Lexile reading metric (Stenner & Fisher Jr., 2013), the corresponding contribution to task difficulty from the "scarcity" frequency, f_k, of word k is the entropy-based term: $\beta_2 \cdot 1/L \cdot \sum_{k=1}^{L} \ln(f_k)$, in a sentence of length L (i.e., the number of words) that is added to the overall CSE for RAVLT task difficulty.

2.4.6 Multidimensionality and PCA: construct alleys – sensitivity and scale distortion

Serial position effects (SPEs) in word learning lists (Section 2.4.5.3) have recently come into focus as potential markers for Alzheimer's disease (AD) and mild cognitive impediment (MCI) (Weitzner & Calamia, 2020). This diagnostic potential of SPE in AVLT implies a potential breakdown in the assumption of specific objectivity of the model (Section 2.3.5.1), which is needed for metrological invariance. This is simply because if SPE have diagnostic potential, then different portions of the cohort – sorted according to health status – will experience different task difficulties.

Green and Smith (1987) give a comprehensive account of the assumptions behind a CSE formulation of Rasch attributes (following Fischer (1973), they denoted the formulation *linear logistic test model* (LLTM)). The extent to which the basic tenet of unidimensionality in the Rasch attribute would be challenged by SPE in RAVLT was studied in a recent Rasch-based analysis (Melin et al., 2022).

2.4.6.1 PCA loading explained

We successfully explained PCA loading in both CSE formulation and in logistic regression residuals in terms of entropy using the Brillouin (1962) formula (eq. (2.18)) in word learning list tests (Melin et al., 2022). Because we can explain task difficulty when formulating CSE, particularly deploying the concept of entropy, it should be possible to identify factors common to both the CSE PCA (as the first step in a PCR – Section 2.3.2) and the PCA of logistic fit residuals. The effects of scale distortion are therefore predictable, for instance, in loading plots as in the present study where different test persons have more or less discrimination to SPEs according to their cognitive state. Empirical evidence was found that SPE represented additional dimensions of task difficulty over and above the difficulty of recalling any word in the RAVLT immediate recall, as evident in the clustering in the PCA loading plot of logistic fit residuals.

Thanks to our MSA approach, models could be formulated of how changes in object properties (such as change in task difficulty, for whatever reason) can propagate through to the subsequent response. An effect of a shift $\Delta\delta_{\mathrm{SPE}}$ in task difficulty, such as that observed in the CSE of different cohort groups due to SPE in word list tests (Melin et al., 2022), will lead in turn to a change, $\Delta P_{\mathrm{success}}$, in response, that is, a change in the probability of making a correct classification (Section 2.3.4.2). In any measurement system, the instrument (B) has a certain sensitivity, $K = \partial P_{\mathrm{success}} / \partial\delta$, that is, how much each test person responds to a task of a certain difficulty, δ. According to an MSA approach (Section 2.3.3), the amount of response change depends on both the magnitude of the change in difficulty as well as the sensitivity K at any given level of task difficulty. A change in task difficulty $\Delta\delta$ for each item j (such as associated with SPE) can lead to a change in logistic fit item residual, $y_{i,j}$, of the regression of the Rasch formula to raw response data where the task sensitivity $K_\delta = \partial P_{\text{success}} / \partial \delta$:

$$y'_{j,\delta} = y_j - \frac{\partial P_{\text{success}}}{\partial \delta} \cdot \Delta\delta \tag{2.19}$$

by a simple first-order partial differentiation.

PCA of logistic fit residuals provides evidence of SPE-related effects as studied in word list memory tests by Melin et al. (2022). In particular, the loading of a PC in the logistic regression residuals will be proportional to the product of the sensitivity (K_δ), and a change in perceived task difficulty, $\Delta\delta$:

$$\text{PCA residual loading: } L_{p,x} \propto a_{p,x} \cdot \frac{\partial P_{\text{success}}}{\partial \delta} \cdot \Delta\delta \tag{2.20}$$

Term for term on the RHS of eq. (2.20):

i. The loading coefficient $a_{p,x}$ should be the same as deduced in the PCA when forming the CSE for task difficulty in terms of the entropy-based explanatory variables (Brillouin's (1962) eq. (2.18)).
ii. $K_\delta = \partial P_{\text{success}} / \partial \delta$, the peculiar **sensitivity** of the instrument (person) in the Rasch model, will "modify" the PCA loading plots correspondingly but can be calculated from a simple differentiation of the dichotomous Rasch formula (Pendrill & Petersson, 2016).
iii. The third term on the RHS of eq. (2.20) is any significant change, $\Delta\delta$, in task difficulty.

Our RAVLT work (Melin et al., 2022) indeed showed some effects of scale distortion that might be correlated with different diagnostic groups, although measurement uncertainties were relatively large (reflecting the limited sample size). We suggested, nevertheless, that (Melin et al., 2022): *what appears to be the case is that, over and above individual variations in a person's ability, there is an overall shift in the person's ability for each clinical group. Whether one regards that as a change in ability or a change in task difficulty is a moot point.*

2.4.6.2 Construct alleys explained

The same MSA model of logistic regression residuals (eq. (2.19)) used to explain PCA loading plots, as modified by a change in stimulus (such as change in task difficulty) in proportion to the instrument sensitivity, can also be used to explain how so-called construct alleys work in revealing scale distortion. For instance, sufferers of onerous diseases such as myotonic dystrophy (DM) type 1 (Hermans et al., 2015), in construct alley plots of task difficulty values, δ (or person ability values, θ) against the residuals, such as INFIT − ZSTD, $\frac{2 \cdot \left(\sqrt[3]{X} - \sqrt[3]{N_{TP} - (2/3)}\right)}{\sigma}; X = y^2$, of the logistic regression (Massof,

2014) appear to scale extrovert and introvert activities differently, producing construct alleys of opposite signs (Pendrill, 2019).

2.4.7 “Vertical” complexity and hierarchical task difficulty: metrological standards

In order to gauge person (or, more generally, agent) ability, normally a set of tasks spanning a range of task difficulty is established.

In attempting to explain a hierarchy of tasks ranked by degree of complexity, Commons et al. (2014) claims that: “task sequences of task behaviors form hierarchies (which) become increasingly complex.” In Commons’ et al. (2014, pp. 9, 10):

> Hierarchical complexity refers to tasks that require the performance of lower-level subtasks in order to perform more complex, higher-level tasks. A higher order action is defined in terms of two or more order actions of one order below, and the higher order action non-arbitrarily organizes those next lower order actions.

Commons et al. (2014) give, as an example, the task of performing

> long multiplication, such as a × (b + c), as organizing the lower order actions of addition and multiplication, in non-arbitrary ways . . . . Orders of hierarchical complexity form an ordinal scale with the first four axioms and definitions that follow. A fifth axiom makes all of the orders of hierarchical complexity equally spaced – that is, of equal difficulty.

This raises the question: why is “complexity” (at least in their terms) the basis for a hierarchy and not some other criterion? Is it really so, as Commons et al. (2014) boldly claim: “there is only one sequence of order of hierarchical complexity of tasks *in all domains*” (our italics)?.

There have been many attempts over the decades to explain the level of difficulty in various mathematical operations; see for example Fischer (1973).

When formulating measurement relations, our recommendation is to use the engineer’s measurement system equation that relates the response of an instrument to a stimulus from the measurement object of a measurement system – (eq. (2.4), equation 2.9 of Bentley (2005)). Our findings (Section 2.4.5) about task difficulty explanations based on entropy-based CSE suggest the following:

- a hierarchy of tasks, ranked in terms of increasing difficulty, arises naturally from the Brillouin (1962) combinatoric approach where the number, G, of categories available for classification increases in each successive test sequence.
- the amount of “useful” information in a message applies not only to elementary bits – binary digits, such 0 and 1 – but also to more sophisticated information content, depending on the “readability” and “information content” of the messages (Table 2.4).

Our explanation of task difficulty – if sufficiently well understood – enables a CSE to be formulated (Section 2.3.2, using Brillouin's (1962) concepts, for instance), which in turn enables tasks or abilities to be ordered hierarchically.

There is a well-known view about "units" on a logit scale (Linacre & Wright, 1989), but which is mainly about a mathematical as opposed to metrological unit. The logit scaling is purely mathematical, but a set of items of known task difficulty (particularly when explained with a CSE) establishes a scale with a definite unit (as well as a natural origin). In choosing a "fit-for-purpose" metrological reference, the attribute of the measurement object (e.g., task difficulty) is the most usual choice of metrological reference (i.e., etalon) in which to realize a unit, in the same way as one chooses a weight as a mass standard (rather than a balance). Indeed, as we have recently proposed (Pendrill, 2019), a CSE for task difficulty enables metrological references in cognitive memory tests analogous to certified reference procedures for metrology in Chemistry, e.g., see the NeuroMET project, https://www.lgcgroup.com/our-programmes/empir-neuromet/neuromet2/. A CSE can provide a recipe for predicting the task difficulty in memory tests in terms of factors – such as the degree of order in the task, thus allowing a calibrated item bank to be established (Melin et al., 2023). Choosing a (calibrated) measurement instrument as a standard (such as a testee) can also be done but is less usual for reasons of practicality.

No attempt is made here (as done by Commons et al., 2014) to make different task difficulties to be equally spaced. Instead, if required, conversion of unfamiliar logits to a 0–100 scale can, of course, be readily done.

Summarizing the use of CSEs in calibrating measurement instruments even in psychometric contexts, in contrast to a common opinion that: "it is extremely difficult to determine what [the instrument] function is for any given attribute/instrument (Kellen et al., 2021)," in the Man as a measurement approach (Section 2.3.1) that function is readily determined in the same way as is done regularly in classical measurement engineering. Indeed, the most recent research on neurodegeneration has, in fact, formulated metrological references for cognitive task difficulty, which can be used to calibrate the measurement system function (Melin et al., 2021).

In addition to enabling metrological references with CSEs, another added value of our work has been significant reductions in measurement uncertainty for a number of legacy memory tests and the establishment of a new and more effective NMM built on a coherent and bespoke selection of individual test items, thanks to implementation of Rasch measurement theory and CSEs (Melin et al., 2023).

These improvements in task difficulty estimation also feed through to corresponding improvements in the determination of the cognitive ability of each test person. That, in turn, enables better correlation studies in which cognitive ability is explained in terms of biomarkers (Melin et al., 2023). Ultimately, even person ability is expected to be explainable in terms of entropy – a more ordered person is usually more able.

2.4.8 Amount of useful information: chunking

Brillouin’s (1962) insight is that it is not simply the number of bits of information, but rather the number of *combinations* of symbols that determines the amount of “useful” information in a message. Importantly, this insight applies not only to elementary bits – binary digits, such 0 and 1 – but also to more sophisticated information content, depending on the “readability” and “information content” of the messages “higher up” the information hierarchy (Table 2.4).

An elementary syntax sequence – such as a set of blocks in the CBT memory test – can be formulated as a “message” in a number of alternative ways, depending on what one considers to a unit of information (“chunk” or “word”). Brillouin’s (1962) formula (eq. (2.18)) applies as earlier in this section, but instead of identifying the number of combinations with simply the number of blocks in a sequence, useful information can also be communicated as patterns (or “figures”) of blocks, such as shown in Figure 2.6, if that is thought as a useful way for a responder when recalling each sequence.

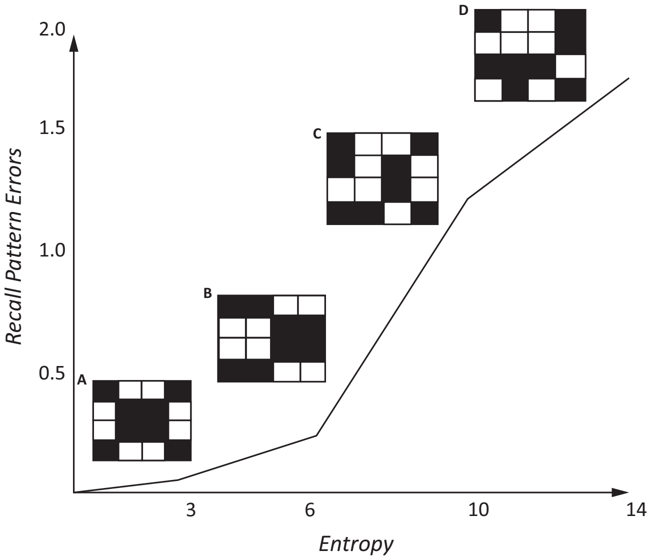


Figure 2.6: Modeling block sequences (adapted from Schnore and Partington, 1967).

Such modeling of the difficulty of remembering block sequences in terms of shapes, figures or patterns rather than the simple number of blocks was performed by Schnore and Partington (1967) who wrote in a description of the pattern shown in Figure 2.6 meant to represent a block sequence:

> For Pattern A, the occurrence of a black or white cell was determined randomly for the four cells in the upper left quadrant, with the constraint that two of the cells had to be black. The remaining

> quadrants of the pattern were obtained by reflecting vertically and then horizontally the quadrant for which the nature of the cells was determined randomly. Thus, Pattern A was symmetrical vertically as well as horizontally with the axes of symmetry passing between the second and third column and row. Type A patterns may be said to contain 2.6 bits of information because only six distinct patterns can be constructed under the rules outlined above: $\frac{4!}{2!\cdot 2!}=6; \ln_2(6)=2,6$ *bits* (eq. (2.17)). The three quadrants of Pattern A which were obtained by reflection may be considered to be redundant and not adding any further uncertainty." Schnore and Partington (1967) found that the number of recall pattern errors amongst 214 university summer school students increased in proportion to the task entropy for a series of patterns with successfully less symmetry. (Figure 2.6)

The approach of Schnore and Partington (1967) to explaining a two-dimensional block recall test in terms of the number of permutations of distinct patterns of blocks is obviously different, but related, to our approach using eq. (2.18). The difference is principally in what one considers to be the "symbols" in Brillouin's (1962) expression. In our approach, a symbol is each block and "*G*" is the number of distinct blocks. Schnore and Partington (1967) consider instead a "symbol" to be a distinct *pattern* of blocks.

Rossi-Arnaud et al. (2006) in their continuation of matrix block studies, commented:

> Research on complexity judgments of matrix patterns has in fact shown that the concept of complexity is determined by both a quantitative and a structural factor. . . . Quantitative complexity includes aspects such as the number of elements in a stimulus and the size of a stimulus. Structural complexity is related to the redundancy of a stimulus. A stimulus is redundant if parts of it can be predicted from other parts. Gestalt factors including symmetry, good continuation, and other forms of regularity constitute redundancy.

Those authors continue by considering the concept of "chunking" (Miller, 1956):

> This is turn raises the question of how the chunks are formed, whether they rely on relatively automatic processes, or are dependent on active manipulation within working memory. It seems likely that both methods of chunking exist. Immediate recall of briefly presented chess positions is substantially greater in expert than in novice players . . ., presumably because the expert can chunk more effectively. However, when given a demanding executive, or visual spatial concurrent task, both experts and novices show impaired performance . . . suggesting that attention and visuo-spatial processing are necessary for this type of chunking. It seems possible however that other more low-level visual components of chunking may be automatic rather than executive. One obvious candidate for this might be symmetry.

Helm (2000) refers to a unit of measurement for the structural information in a message as a "sip" (structural information parameter) corresponding to one structural degree of freedom (at no matter which hierarchical level). For each degree of metrical freedom (i.e., at the lowest hierarchical level in a code), the assumed resolution implies that λ metrical variations are to be distinguished. If these metrical variations can be specified by the decimal numerals 1, 2, 3, . . ., λ, then the binary specification of one of these metrical variations requires $\log_2(\lambda)$ bits. This implies that sips at the lowest hierarchical level in a code can be converted into bits by means of the equation:

$$1\,\text{sip} = \log_2(\lambda)\ \text{bit}$$

Helm (2000) continues: the "precisal" reflects a set-based property (i.e., a probability) of an object, defined in terms of an object-based property (i.e., a complexity) of the object. This contrasts with the surprisal, which, inversely, reflects an object-based property defined in terms of a set-based property.

From our point of view, a sip constitutes a "word" in the most general sense. An example is the explanation of task difficulty in recalling a certain list of symbols (G blocks, digits, words, etc.) as the "object" of the measurement system, where the λ (metrical) variations are the different possible combinations $(G!)$ of symbols and the amount of information (which determines the task difficulty, see eq. (2.16)) is the entropy $-K \cdot \log(G!)$. As in the study of recall difficulty for block sequences by Schnore and Partington (1967), instead of blocks, one can choose patterns based on sequences of different symmetry, if one believes that information is predominantly conveyed in those symmetric patterns instead of blocks.

According to Helm (2000), the intuitive Gestalt notion of "goodness" was generally operationalized empirically in the 1950s and 1960s in terms of, for example, matching, remembering, and learning paradigms. In attempting to explain the human interpretation of visual stimuli (as an example of a measurement process), van der Helm (2000) recalled that the likelihood principle states that the visual system has a preference for the most likely interpretation (i.e., the one with the highest probability of being correct). In contrast, the simplicity principle states that the visual system has a preference for the simplest interpretation (i.e., the one with the shortest description – the least complex). Helm (2000) argues that these two principles – likelihood and simplicity – though similar, are nevertheless distinct and illustrates this by giving the following two equations when perceiving (probability distribution Q, interpretation "H" that equates to the perceived entropy, $H(Q)$) a stimulus (probability distribution P, stimulus "D" that equates to the stimulus entropy, $H(P)$):

- Likelihood principle:
  Select the $H(Q)$ that maximizes $p\{H(Q)|H(P)\} = p(H(Q)) \cdot p\{H(P)|H(Q)\}$.
- Simplicity principle:
  Select the $H(Q)$ that minimizes $I\{H(Q)|H(P)\} = I(H(Q) + I\{H(P)|H(Q)\}$.

These two equations become equivalent when the information $I = -\log_2(p)$ (probabilistic "surprisal") or equivalently $p = 2^{-I}$ (descriptive "precisal") (Helm, 2000).

We consider symmetry further in the next section. The choice of "symbol" in different cases reflects which element in a message is considered in each case to be the bearer of information, that is, the "word." A similar consideration can be made when extending our approach to a wide range of applications, such as identifying which elements of an artwork have some kind of meaning when appreciating the work (Sigaki et al., 2018), as explored further in Section 2.5.

2.4.9 Symmetry, conserved quantities, and minimized entropy

The choice of systems suitable for units (Section 2.2.4) is not only about the number of decimal places with which the quantity can be measured. Above all, what is crucial is the amount of information communicated in a message, for example, units – as recognizable "packages of measurement information" – that can be measured in terms of entropy (Weaver, 1949; Shannon, 1949).

Languages used range from the simplest syntax (communication of signs or numerical values) via semantic (meaning) and pragmatic messages to fully effective messages that lead to actual changes (Table 2.4).

A central factor is then symmetry – invariance under displacement – a prerequisite for forming "meaningful" words and units of measurement. The more symmetrical – or ordered – a symbol or word is, the lower the entropy.

This is also the case for the different units of measurement, where one sees recognizable patterns that reflect transformational symmetry (low entropy). For example, the unit of time (t) is found in various physical systems (a clock, atomic transition, planet, pulsar . . ., as in the SI definition of the second), where in all cases the canonical variable is energy (E) as preserved during a "displacement" through time. Likewise, a ruler for measuring length (l) (or measure for rotation) is assumed not to change its length when a certain distance is measured by displacement, thanks to Lorentz invariance associated with conservation of momentum (p). It is the same displacement that Maxwell addresses in his text on units (Section 2.2.4).

It is well known that transformation symmetry is related to quantity conservation, which is precisely what is sought when defining units of *quantity*. When looking for suitable systems to define and realize units, one can observe that a number of physical quantities are known through experiment to be conserved in isolated systems: the total energy, momentum, and angular momentum remain constant, regardless and however complex interactions occur in the system.

These constants are, in turn, consequences of invariance in mechanical systems under changes of the corresponding canonical quantities – time, displacement in length and during rotation in space – together with the principle of least action (Dirac, 1992). Concepts such as entropy and symmetry in measurement are useful, not only in physics – where invariance is a fundamental characteristic – but also in a number of applications, where units are now sought for equity in, for example, the human and social sciences.

To capture in full generality this concept of efficiently carrying an amount of information, recourse can be made to similarity transformations in matrix algebra. An arbitrary representation **v** (or pattern or message) can in general be decomposed into subsets if a similar (or displacement) matrix **D** can be found for the similarity transformation $\mathbf{v}'(a) = \mathbf{D}^{-1} \cdot \mathbf{v}(a) \cdot \mathbf{D}$, which diagonalizes every matrix in the representation into the same pattern of diagonal blocks – each of the blocks is a representation of the group independent of each other. When no further decomposition becomes possible, the representation is said to be irreducible.[3]

2.4.10 Quantum mechanics and measurement

Measurement is often mentioned when describing quantum mechanics, often with reference to the impossibility – due to the finite value of the Planck constant – to make a measurement without disturbing the object being measured. This includes the famous Heisenberg uncertainty ratio as well as many topics in contemporary physics, such as quantum entanglement and the possibilities of making quantum computers.

Another aspect that we want to highlight is how the measurement process is described in quantum mechanics, as a template for a description of measurement more generally. Eigenstates naturally play a well-known role in quantum mechanics, and when introducing the expression: $\mathbf{Q}|q\rangle = q|q\rangle$ for a quantity **Q**, Dirac (1992, section 10, p. 35) mentions the measurement aspects: "If the dynamical system is in an eigenstate of a real, dynamical variable **Q**, belonging to the eigenvalue *q*, a *measurement* [my italics] of **Q** will certainly yield the result *q*." It is worthwhile considering the similarities between this quantum-mechanical expression and the corresponding measurement engineering expression for the response, *O*, of a measurement system to a stimulus, *I*, as given by eq. (2.4).

In line with our discussion of units in connection with symmetry and entropy (Section 2.4.9), we would like to go beyond considering a physical constant only as a "simple number." Remember Maxwell's classic words in connection with eq. (2.1) that "make up" a certain amount and count how many times a unit fits into the measured displacement, where "displacement" is not specifically in length, but in the dimension of interest. The measure of displacement can be interpreted in the broadest sense, not only in the physicist's laboratory but more generally include the scalability required in all applications.

A connection between Dirac's and Maxwell's descriptions of displaced measurement systems (Pendrill, 2019) can be established by observing that the unit quantity has an eigenvalue q_{unit}: $[Q]|q\rangle = \{q_{\text{unit}}\}|q\rangle$, and a displaced observable $Q^*|q\rangle = \{Q\} \cdot [Q]|q\rangle = \{Q\} \cdot \{q_{\text{unit}}\}|q\rangle$, where Dirac's "unitary" operator:

$$U = \{q_{\text{unit}}\} \cdot [Q]$$

The quantization rule: $\oint p \cdot dq = n \cdot h$ for a pair of canonical variables (*p, q*) such as position and momentum or time and energy (Born, 1972), means that the integral in one period of the motion gives an area that is an integral multiple of *h*, according to the quantum postulate. The eigenfunctions of the square of the angular momentum of the

[3] https://en.wikipedia.org/wiki/Irreducible_representation.

operator, as irreducible representations, form a set of base functions accompanied by a set of eigenvalues (multiples of h) that can be considered as units of measure, which can appear in various physical systems suitable for defining units of measure: for example, vortices in superconducting quantized Hall effect system. For our purposes in metrology, it illustrates how the Planck constant, h, functions as a fundamental unit of measurement. The Boltzmann constant, k, also work in a similar way, such as when describing a quantized amount of information in terms of entropy: $-k \cdot \ln(2)$ is the smallest communicable information when a "bit" (multiplicity = 2) to be conveyed (Cohen-Tannoudji, 2009).

Of relevance to our discussion of fundamental constants as units (Section 2.4.9), Dirac (1992) also briefly mentions the case when a dynamical variable is a number – then each state becomes an eigenstate and the dynamical variable is obviously observable. Every measurement of it always gives the same result, so it is "just" a physical constant, like the charge on an electron. A physical constant in quantum mechanics can thus be seen as:

- either as a quantity with a single eigenvalue
- or as a simple number shown in the equations,

where these two viewpoints, according to Dirac, are equivalent.

Again, descriptions of how the fundamental physical constants are now included in definitions of units of measurement are found in the new SI, 9th edition, SI Booklet (CGPM, 2018a, b). According to our description, one can search for units of measure more generally among fundamental symmetries described in terms of minimum entropy.

2.5 Linguistic, artistic, and metrological standards and meaningful communication

In this final section, the methodology of the previous sections will be extended from the simplest syntax to a number of applications of increasingly elaborate information content (Table 2.3), including, for instance, literature (Sections 2.5.1 and 2.5.2), urban scenes (Section 2.5.3), and fine art (Section 2.5.4). In many cases, connections are made between entropy and task difficulty, as in Section 2.4.5.

But – except where noted – in most cases the Rasch psychometric approach (Section 2.3.5) does not yet seem to have been deployed. In such cases, metrological quality assurance will be difficult since limitations – an inherent nonlinearity, effects of ordinality and a general confounding of task difficulty and instrument ability (Table 2.2) – are expected to lead to unnecessarily large uncertainties, with substantial risks of incorrect decisions and potentially serious consequences in many fields of application.

2.5.1 Communication and probabilistic language modeling

The task difficulty associated with understanding a language sentence can be explained at various levels of sophistication in a way analogous to our developing view of task difficulty in syntax-based tasks such as elementary memory recall tests (Section 2.4.5) where, in addition to the number of combinations of blocks, one also considered the number of distinct patterns or figures formed by the blocks in each test sequence.

Language complexity during comprehension can be accounted for computationally explicitly in probabilistic language models, as reviewed recently by Armenia et al. (2017) in the context of developing neurobiological models. Statistical or probabilistic language models assign conditional probabilities to linguistic representations (e.g., words, words' parts-of-speech, or syntactic structures) in a sequence. Together with information-theoretic complexity measures, estimates can be made of word-by-word comprehension difficulty in neuroscience studies of language comprehension.

A common application of probabilistic language modeling is the task of sequence-prediction, where expectations can be generated about upcoming words given the words seen so far in a sentence. A distinction can be made between "statistical language modeling," that is, predicting the words based on sequences of past words, and models that estimate the probability of a syntactic structure underlying the observed sequence of words or the probability of the upcoming word given the syntactic parse so far (Armenia et al., 2017). In either case, the uncertainty in the prediction is, of course, a kind of decision quandary that can be analyzed with the Rasch psychometric model (Section 2.3.5) where the probability of a correct classification is expressed in terms of the ability of the classifier and the level of difficulty of each classification task (Bashkansky & Turetsky, 2016; Pendrill et al., 2023).

One of the simplest architectures for estimating probabilities, namely the *n*-gram — as used extensively in reading metrics (Stenner & Fisher, 2013) — is described in the recent review of probabilistic language models by Armenia et al. (2017). An *n*-gram model takes into account the (n – 1) preceding words in a sequence for computing the conditional probability of occurrence for the *n*th word based on the relative frequencies of co-occurrence of word sequences derived from the training data in language corpora. An *n*-gram can stand for the sequence of actual words or, alternatively, syntactic categories of words (or parts-of-speech).

Quantifying complexity, entropy, and surprisal is also described in the recent review of probabilistic language models by Armenia et al. (2017). An early example of indices to capture patterns of symbols is the work of Orlitsky et al. (2006). An index $\iota_{\bar{x}}(x)$ of x is one more than number of distinct symbols preceding x's first appearance in the sequence $\bar{x} = x_1, \ldots, x_n$. The "pattern" of $\bar{x}$ is then:

$$\psi(\bar{x}) \overset{def}{=} \iota_{\bar{x}}(x_1) \cdot \iota_{\bar{x}}(x_2) \cdot \iota_{\bar{x}}(x_2) \cdot \cdots \cdot \iota_{\bar{x}}(x_n)$$

and the probability sequence generated according to p (distribution of x) has pattern $\bar{\psi}$

$$p(\bar{\psi}) \overset{def}{=} p(\{\bar{x}:\psi(\bar{x}) = \bar{\psi}\})$$

Using the Shannon expression (eq. (2.9)), entropies of sequence $\bar{X} = X_1, \ldots, X_n$ with its pattern $\bar{\psi} = \psi_1, \ldots, \psi_n$ are given by Orlitsky et al. (2006) as

$$H(\bar{X}) = -\sum_{\bar{x}} p(\bar{x}) \cdot \log(p(\bar{x})) \qquad H(\bar{\psi}) = -\sum_{\bar{\psi}} p(\bar{\psi}) \cdot \log(p(\bar{\psi}))$$

Helm (2000) also considers patterns and describes interpretations of visual stimuli by means of certain coding rules that each “squeeze out” a specific kind of regularity. For visual perception the following three rules allow for “hierarchically transparent” descriptions of “holographic regularity,” that is, the structure of regularity should be such that all its substructures reflect the same kind of regularity, or, equivalently stated, the regularity should be invariant under growth (as opposed to the transformational approach which generally focuses on invariance under motion):

Iteration:	$kkk \cdots kk$	Code:	$m*(k)$
Symmetry:	$k_1\, k_2\, \cdots k_s\; p\; k_s \cdots k_2\, k_1$	Code:	$S[(k_1)(k_2) \cdots (k_s),\ (p)]$
Alternation:	$k\ x_1 k\ x_2 \cdots k\ x_n$	Code:	$\langle (k) \rangle / \langle (x_1)(x_2) \cdots (x_n) \rangle$
	$x_1 k x_2 k \cdots x_n k$	Code:	$\langle (x_1)(x_2) \cdots (x_n) \rangle / \langle (k) \rangle$

A complexity measure of a message (Helm, 2000) includes degrees of freedom at different hierarchical levels and is not merely syntactical but semantic in that it derives from a perceptually meaningful classification of patterns. For instance, for the message *abababababab*, the code $2*(3*(ab))$ has three hierarchical levels: $2*(X)$ where X reflects a degree of freedom; the second level is $3*(Y)$ where Y reflects a degree of freedom; and the third level is *ab* where a and b each reflect a degree of (metrical) freedom – hence, four degrees of freedom in total.

2.5.2 Probabilistic language models in cognitive neuroscience: Summarization of datasets. Meaning and importance – Knowledge and cognition

The original theory of Shannon (eq. (2.9)), although dealing primarily with syntax (i.e., the lowest level in the information hierarchy shown in Table 2.4), can operate at the semantic level by relying on semantic units (Weaver, 1949).

Identification of the most important information in the context of dataset analysis from a source to produce a comprehensive output for a particular user and task is termed “summarization” (Peyrard, 2019). For summarization, a text X can be considered (Peyrard, 2019) as a “source emitting semantic units . . . represented by a probability distribution $\mathbb{P}_X$ over the set Ω of semantic units.” Importance arises according to that author as a “single quantity naturally unifying three concepts: Redundancy, Relevance and Informativeness.” He recalls different interpretations and motivations for these concepts.

More in line with determining the difficulty of interpreting the source document, the more recent work of Khurana and Bhatnagar (2022) on summarization focusses in more depth on terms, sentences, and topics (in latent space) as the three semantic units when “capturing the essence of a document.” Wells (2011) considers terminology and concepts in this area in more detail.

In a recent study of entropy and Bloom’s taxonomy in an educational context, Larsen et al. (2022) “empirically examine two major assumptions:

– the independence of the knowledge-type and cognitive-process dimensions:

Factual and *conceptual* knowledge can be distinguished based on the context of the question.

– the use of action verbs as proxies for different cognitive processes – in hierarchical order: remembering, understanding, analysing, applying, evaluating and creating”:

Knowledge-type and cognitive-process dimensions were found to be related and not independent with two principal axes in how the two dimensions are related, with three clusters of knowledge types and cognitive processes. Entropy was also considered by Larsen et al. (2022) when:

> using the Shannon evenness index: $J' = H'/\ln(L)$, where $H' = \sum_k^L p_k \cdot \ln(p_k)$ where L is the number of categories or cognitive processes observed, p_k the proportion of a specific cognitive process used out of the total frequency of a given prompt word, and k the index for the different cognitive processes. They did not find a clear relationship between question prompt words (including action verbs) and cognitive processes in the assessment items.

2.5.3 Environmental stress

That long-term exposure to stress can lead to chronic effects will be no surprise and is readily mentioned in the literature on stress, for example, Pretty et al. (2005). Here we give two examples of studies: first, measures of accessibility perceived during train journeys (Section 2.5.3.1) and then stress in the built environment (Section 2.5.3.2).

2.5.3.1 Accessibility when making a journey

A mathematical model for the aggregate accessibility (A_{ij}) (Berglund et al., 2014) for a person (i) making a complete journey (j) – from the initial planning, through travel, until arriving at the final destination (the whole trip) – is

$$A_{ij}^{m} = \prod_{b} \left[1 - p_{jb}\psi_i(d_b)\right]$$

as a product over the series of barriers, b, that have to be overcome during the journey of the individual person whose perceived effort ("hinder" or "level of hindrance," see Table 2.2) is ψ_i when facing a certain barrier at a distance d_b, together with the probability, p_{jb}, that a person, i, will face that particular barrier (Church & Marston, 2003).

Berglund et al. (2014) made an extensive survey of travelers' experiences of train journeys in the greater Stockholm area. The *measured* perceived effort ψ_i for a person when encountering a barrier, b ($p_b = 1$) during a train journey, will be a function of the true value, φ, and an error component ε: A linear measurement model can be adopted so that the relationship between the measured perceived effort for each barrier and the true value can be written as

$$\psi_i(d_b) = \varphi_i(d_b) + \varepsilon_{ib}.$$

The error component, ε, is a random variable that we assume is normally distributed with a mean value, $\bar{\varepsilon}$, and variance, σ_{ε}^{2}.

The perceived effort function, ψ_i is defined between 0% and 100%, and a key observation is that it can be converted into an accessibility score:

$$P_{\text{success}} = 100\% - \psi$$

of how successful deployment of the system is for each task at hand (i.e., the service of providing good rail travel). Thus, the function value $\psi = 100\%$ indicates a barrier such that the probability of cancelling the journey is 100% when facing such a barrier. A value of "0" indicates the opposite, that is, the barrier causes no problem to the traveler and accessibility is complete, that is, $P_{\text{success}} = 100\%$. The accessibility score for each individual traveler when negotiating a barrier depends on task challenge, δ, and traveler capability, θ, according to the Rasch psychometric model (eq. (2.8), Section 2.3.5).

The Rasch model was demonstrated in this project (Berglund et al., 2014) to be an efficient tool for analyzing how both person attributes and item attributes contribute to the overarching concept of accessibility in train traveling. Among the results of this study, individual person attribute values indicated a considerable spread, skewed away from the corresponding distribution of the measured item attributes that were heavily skewed toward the more capable test persons (TP), reflecting the spread of functional limitations of the persons studied. For the items, broadly speaking, questions concerning ergonomic barriers (such as "to get off the train") appear on the average to indicate less challenge than questions concerning informational (or cognitive) barriers (such as "to retrieve information on-board").

2.5.3.2 Neurophysiological responses to the built environment

Guidi et al. (2021) have recently made a systematic review of allostatic load and its impact on health, where allostatic load is explicitly "the cumulative burden of chronic stress and life events":

- Direct, overt, and easily quantifiable factors that might affect health and wellness when considering the impact of architecture include light exposure, noise levels, air pollution, and ambient temperature (Cedeño-Laurent et al., 2018).
- Visual exposure to certain subtle variations in the shape or configuration of the built environment, "architectural forms," such as room width and wall curvature, can elicit neuroimmunological stress responses (Shemesh et al., 2021).

Stress response is not merely registered with clinical biomarkers. Indeed, stress responses can be both **affective, behavioral, or biological** (Cohen et al., 1995). Recent stress research employing Rasch measurement theory is the work of Hadžibajramović et al. (2015).

Considering first the propensity of a particular scene to induce stress, in Rasch terms (Section 2.3.5): $\widehat{\boldsymbol{Y}} = \delta$, task "difficulty," propensity to stress that could be added to the examples given in Table 2.2.

Fernandez and Wilkins (2008) found their findings consistent with the idea that the visual system has evolved to process natural scenes efficiently and that images with unnatural statistics can sometimes be stressful physiologically, particularly when they have an excess of energy at spatial frequencies to which the visual system is generally most sensitive. They include an example of art (Jesmond Barn by Debbie Ayles), inspired by an attack of basilar artery migraine. Fernandez and Wilkins (2008) suggested that the statistics of uncomfortable images revealed in their study could be used to avoid the controversies that result when stressful images appear in contexts where such images are inappropriate, as, for example, in public art, particularly art in hospitals. In Section 2.5.4, we will consider artwork characterization in terms of structural entropy and complexity, which has some similarities with characterizing the built environment.

Secondly, considering the sensitivity (or "discrimination") of individuals to allostatic load, in Rasch terms (Section 2.3.5): $\widehat{\boldsymbol{Y}} = \theta$, sensitivity of a person to become stressed (or, equivalently, each individual's resistance to stress). Johnson et al. (2019) assumed that the observed variables are indicators of the unobserved latent variables and estimated a model that specifies the number of latent profiles or classes, as well as the relationships between the latent variables and the observed variables:

- Four **latent physiological risk** profiles: low, metabolic, inflammatory, hypertension
- Ten **biomarkers, $\mathbf{X}_k$**: diastolic blood pressure (DBP), systolic blood pressure (SBP), pulse (PLS), C-reactive protein (CRP), glycohemoglobin (GLY), albumin (ALB), creatinine clearance (CREAT), Body Mass Index (BMI), high-density lipoprotein (HDL), total cholesterol (CHO)

Latent profile analysis (LPA) is a model-based approach that has been applied recently when measuring allostatic load in clinimetric research – that is, where clinical biomarkers are complemented by incorporating patient-reported symptoms and physical signs of health outcomes (Fava et al., 2022).

Comparisons should be able to be made with the formulation of CSEs for person ability in memory tests as functions of various biomarkers (Section 2.4.7).

2.5.4 Artwork characterization in terms of structural entropy and complexity

The characteristics of an artwork can be quantified and compared by considering characteristics such as shape, colors, and brightness, which have some kind of meaning when appreciating the artwork.

Work in the field of digital image processing and recent advances in finding automated measures of visual complexity has, according to Marin and Leder (2013), profited the study of subjective complexity and its relations to aesthetic experience. They studied the effects of the number and variety of elements present in various visual and auditory scenes that were found to be the strongest determinants of subjective complexity, more than organization or symmetry (Nadal et al., 2010; Berlyne et al., 1968).

2.5.4.1 Entropy and the production process

According to Bense (1969) when interpreting Birkhoff's (1933) *aesthetic* measure:

> In any artistic process of creation, we have a determined repertoire of elements (such as a palette of colours, sounds, phonemes, etc.) which is "transmitted" to the final product. The creative process is a selective process (i.e., to create is to select). For instance, if the repertoire is given by a palette of colours with a probability distribution, the final product (a painting) is a selection (a realization) of this palette on a canvas. In general, in an artistic process, **order** is produced from disorder. The distribution of elements of an aesthetic state has a certain order and the repertoire shows a certain **complexity**.

Procedures to create artwork algorithmically belong to production rather than characterizing the artwork in itself or subsequent measurement. As summarized by Rigau et al. (2007), Nake (1974), one of the pioneers of the *computer or algorithmic art* (i.e., art explicitly generated by an algorithm), considered a painting as a hierarchy of signs, where at each level of the hierarchy the statistical information content could be determined. He conceived the computer as a *universal picture generator* capable of "creating every possible picture out of a combination of available picture elements and colours." A modern example is the work of Prisma Labs (2023).

Our view is that Bense's (1969) measure – although he referred to it as an informational measure and to a repertoire being "transmitted" – is conceptually distinct from two later stages in appreciating artwork, namely:

(i) a characterization of the artwork in itself (i.e., after production) and

(ii) the aesthetic value of the artwork as perceived by a third party, in a process similar to measurement where the human observer acts as a measurement instrument.

The fact that sets of similar concepts can be deployed at each of these three stages – production, product per se, and perception – does not mean that these stages are equivalent.

2.5.4.2 Visual order of artworks in terms of entropy

As summarized by Rigau et al. (2007), Bense (1969) transformed Birkhoff's (1933) *aesthetic* measure into an informational measure: redundancy divided by statistical information (entropy). Expressions of image order were given by Rigau et al. (2007) in terms of the Kolmogorov complexity and what they claimed was a "new" aesthetic measure $M_S = \left(NH_p - K\right)/NH_p$, where NH_p is the information content of an image and K is the Kolmogorov complexity, using Zurek's (1989) concept of physical entropy. Their measure is the ratio between the reduction of uncertainty (due to the compression achieved) and the initial information content of the image. This was based on the work of Bense (1969) and Moles (1968) who proposed to measure the order in an aesthetic object in terms of redundancy, that is, the "reduction of uncertainty."

Owing to the noncomputability of K, real-world compressors, such as *jpg* and *png*, were used to estimate complexity as it entered the aesthetic measure. Expressions similar to image order given by Rigau et al. (2007), including the ratio of file sizes between original and compressed files, were subsequently applied to marine charts, radar images, icons, line drawings, environmental scenes, and a wide range of artistic works (Forsythe et al., 2011).

Simple "Physics-inspired" metrics that are estimated from local spatial ordering patterns, d, in paintings encode crucial information about the artwork. Having the probability distribution $P = \{p_i;\ i = 1, \ldots, n\}$, Sigaki et al. (2018) calculate the normalized Shannon entropy (eq. (2.9)):

$$H(P) = -\frac{1}{\ln(n)} \cdot \sum_{i=0}^{n} p_i \cdot \ln(p_i), \tag{2.21}$$

where $n = \left(d_x \cdot d_y\right)!$ is the number of possible permutations. $\ln(n)$ corresponds to the maximum value of the Shannon entropy $S(P) = -\sum_{i=0}^{n} p_i \cdot \ln(p_i)$, that is, when all permutations are equally likely to occur ($p_i = 1/n$). The value of H quantifies the degree of "disorder" in the occurrence of the pixels of an image. We have $H \approx 1$ if the pixels appear in random order, and $H \approx 0$ if they always appear in the same order.

Among a range of calculations to characterize different art pictures, Kim et al. (2014) included the entropy of a grey-scale image, again using the Shannon expression (eq. (2.9)).

Sigaki et al. (2018):

> The value of H quantifies the degree of disorder in the pixel arrangement of an image: Values close to 1 indicate that pixels appear at random, while values close to zero indicate that pixels appear almost always in the same order. More-regular images (such as those produced by Minimalism) are expected to have small entropy values, while images exhibiting less regularity (such as Pollock's drip paintings) are characterized by large values of entropy.

> The notions of **order/simplicity versus disorder/complexity** (C) in the pixel arrangements of images captured by the complexity–entropy plane partially encode these concepts. Images formed by distinct and outlined parts yield many repetitions of a few ordinal patterns, and, consequently, linear/haptic artworks are described by small values of H and large values of C. On the other hand, images composed of interrelated parts delimited by smudged edges produce more random patterns, and, accordingly, painterly/optic artworks are expected to yield larger values of H and smaller values of C.

Corresponding studies of artistic judgment aptitude of, for instance, laypersons have included the work of Bezruczko et al. (1990, 2016) who have even done Rasch analyses, as a complement to previous neuroaesthetic studies of aesthetic appreciation and sensitivity.

2.5.4.3 Order, "words," and entropy in the visual arts

Kim et al. (2014) report that digital image processing techniques can be used to investigate three quantitative measures of images – the usage of individual colors, the variety of colors, and the roughness of the brightness. They considered color to be like a "word" for a painter and found a difference in color usage between classical paintings and photographs and a significantly low color variety of the medieval period. They investigated how many different kinds of color appear in a painting and how often a certain color is painted, which is similar to Zipf's (1949) plot for word frequencies in literature. It is named as "chromo-spectroscopy."

Similar considerations of the general concept of a "word" (Section 2.4.2) can be made when identifying which elements of an artwork have some kind of meaning when appreciating the work (Sigaki et al. 2018). As in, for instance, reading metrics such as the Lexile, one could imagine that a factor such as how often one encounters a particular "trademark" of an artist – e.g., a Magritte's angel – would reduce the entropy of appreciating the work.

Apart from merely recognizing such a word, a further perhaps more arousing experience is realizing its familiarity. Associating a word with a particular reminiscence with either a positive or negative affect will mean that the change in entropy could be weighted with an impact factor. Yonelinas et al. (2010) in examining two separate memory retrieval processes – **recollection and familiarity** – give the following citation from William James [*The Principles of Psychology* (p. 658)]:

> I enter a friend's room and see on the wall a painting. At first, I have the strange, wondering consciousness, "surely I have seen that before," but when or how does not become clear. There only clings to the picture a sort of penumbra of familiarity, – when suddenly I exclaim: "I have it, it is a copy of part of one of the Fra Angelicos in the Florentine Academy – I recollect it there!"

Suddenly recollecting where one has seen a familiar actor's face before, despite the different role being played, is an experience of more import than a vague familiarity. Yonelinas et al. (2010) state:

> Recollection reflects the retrieval of qualitative information about a specific study episode, such as when or where an event took place, whereas familiarity reflects a more global measure of memory strength or stimulus recency . . . . the hippocampus is particularly important in forming and retrieving the arbitrary associations that support recollection, whereas familiarity depends on regions outside the hippocampus and reflects a by-product of repeated neural processing.

2.5.4.4 Alternative tests for dementia: measuring Man and deep learning

Motivation for developing alternatives to traditional legacy tests of dementia to those presented earlier (Section 2.4.5) can be found in the recent work of Zhu et al. (2021) on "expressive language impairment": "*Screening measures, neuropsychological assessments, and neuroimaging scans are not pragmatic, cost-, or time-efficient approaches for widespread use.*"

Zhu et al. (2021) write: "expressive language impairment is common in AD, such as reduced verbal fluency and syntactic complexity, increased semantic and lexical errors, generating more high-frequency words and shorter utterances, and abnormalities in semantic content."

In relating this and similar approaches to dementia detection as alternatives to more traditional tests, we recommend making a measurement system analysis when describing the actual setup used (Section 2.3.3): that includes identifying what is the measurement object, instrument, and operator and what are the important characteristics attributed to each element: (i) detecting signs of dementia *from language expression by a person* is not the same as (ii) judging the cognitive ability of a person to accurately interpret language as a measure of dementia, such as with the Lexile® reading metric. Similarly, the work of Tanaka et al. (2019) attempts to (i) detect signs of dementia *from facial expressions* – measuring Man, as the measurement object – which is not the same as (ii) judging the cognitive ability of a person to accurately describe a face (as an example of a picture) – Man as a measurement instrument – as is done in traditional legacy memory tests.

Whatever the measurement setup, the probability of success in performing a classification will depend on both the level of difficulty associated with the task (measurement object) as well as the ability of each agent (person, machine acting as an instrument) attempting the classification according to a measurement system analysis (Section 2.3.3).

Modern psychometric methodology (Section 2.3.5) can be deployed to several kinds of classification of potential interest here: (i) dementia detection based on inspection of facial expressions – measuring Man, as the measurement object; (ii) ability of a classifier as a measurement instrument to observe facial expression; and (iii) ROC (Pendrill et al. 2023).

The latest psychometric methods, such as Rasch measurement theory, can provide separate and accurate estimates of task and agent attributes (Hughes et al., 2003).

Innovative dementia detection techniques include deep transfer learning, which according to Zhu et al. (2021) "focuses on storing knowledge gained from an easy-to-obtain large-sized dataset from a general task and applying the knowledge to a downstream task where the downstream data is limited." Such methods need calibration by accessing previous cognitive diagnoses and in earlier work, legacy tests have been used as a "gold standard." For example, from Zhu et al. (2021): "We applied a multitask transfer learning to output both the AD/non-AD labels and the Mini-Mental State Examination (MMSE) scores (a test assessing global cognitive functioning)."

Those same authors found, however, indications of inconsistencies in MMSE: "For regression (of scores against MMSE), RMSE increased from 4.15 to 4.96, which reveals a negative impact of the joint training. This may have been due to the inconsistent cases of MMSE scores and AD/non-AD labels, and the MMSE regression task is more fined-grained and thus received a stronger impact from the inconsistent cases."

Thanks to the latest psychometric methods (Section 2.3.5), it has been known for some time (although not widely recognized) that many memory legacy tests (and other classification tasks) have scale distortion, which may significantly compromise accuracy – see, for example, MMSE (Hughes et al., 2003). If these above-mentioned inconsistencies and shortcomings of for example MMSE as a gold standard when judging diagnostic performance are a serious limitation on validity, then this should be clearly stated.

2.5.5 Measuring aesthetic value

Subjects in the Iosa et al. (2022) study were verbally asked to judge on a numerical rating scale ranging from 0 (not at all) to 10 (the maximum possible) the following psychophysical aspects: how objectively beautiful the picture (exemplified in Figure 2.7) was (objective beauty); how much they liked the stimulus (subjective beauty); and how tiring the exercise was (perceived fatigue). At the same time, the kinematics of each test person's hand movement (shown as red lines in Figure 2.7) when virtually sketching each image included characterization of the entropy is an estimation of the complexity of the trajectory. Entropy can be considered to be inversely related to the efficacy of movements performed to complete the task (similar to Cordier et al., 1994). Iosa et al. (2022) found that: “lower entropy (and hence a less complex trajectory) was associated with paintings than with photos . . . No other variables significantly influenced the complexity of the trajectory . . .. Entropy was significantly correlated with the fatigue perceived for beautiful photos.”

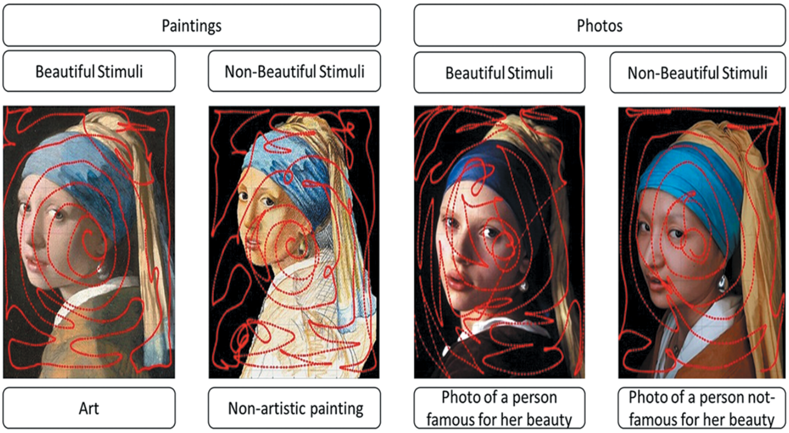


Figure 2.7: Stimuli and responses to paintings and photographs (Iosa et al., 2022).[4]

It is essential to consider the critical role of emotions when accounting for the impact of complexity in aesthetic experiences if one is to avoid a restricted, ecologically invalid study, according to Marin and Leder (2013). Berlyne’s (1971) model predicts that people will generally prefer stimuli of intermediate complexity compared with simple or highly complex stimuli under normal arousal conditions. At the same, Marin and Leder (2013) claim that Berlyne disregarded the “dawn of cognitive psychology.”

After earlier studies of Boon (2011) of how the concept of objective complexity can be associated with artistic paintings and music, Marin and Leder (2013) took the concept of complexity further when considering hedonic measures of preference, pleasantness, and beauty of affective environmental scenes, paintings, and music. In psychology, studies have been made of the behavioral outcomes of sensory, cognitive, and affective responses to stimuli varying in perceived complexity. In empirical aesthetics, one has attempted to understand the impact of stimulus dimensions – such as complexity, uncertainty, and beauty – on hedonic values. In summarizing the results of their experiments, Marin and Leder (2013) found:

- None of the four compression file formats to correlate significantly with subjective complexity judgments of representational paintings. An explanation could be that, for example, some backgrounds in paintings of simple figure-ground compo-

[4] CC BY 4.0.

sitions may consist of a large number of individual brush strokes that form a rather uniform background from a subjective perspective.

- Correlation strength between edge detection measures and subjective complexity is stronger for photographs of environmental scenes than for paintings.

Marin and Leder (2013) claimed convincing evidence for a positive association between subjective complexity and arousal, in line with Berlyne's (1971) collative motivation mode. Complexity correlated moderately positively with arousal ($r_s = 0.36$), in other words, more complex pictures induced higher degrees of arousal.

We finish this chapter by quoting one of the earliest work on aesthetics. Birkhoff (1933) formalized the notion of beauty by the introduction of an *aesthetic measure*, defined as the ratio between *order* and *complexity*, where "the complexity is roughly the number of elements that the image consists of and the order is a measure for the number of regularities found in the image" (Scha & Bod, 1993), as summarized by Rigau et al. (2007).

According to Birkhoff, the aesthetic experience is based on three successive phases:

1. A preliminary effort of attention, which is necessary for the act of perception, and that increases proportionally to the *complexity* (*C*) of the object.
2. The feeling of value or *aesthetic measure* (*M*) that rewards this effort.
3. The verification that the object is characterized by certain harmony, symmetry, or *order* (*O*), which seems to be necessary for the aesthetic effect.

From this analysis of the aesthetic experience, Birkhoff (1933) suggested that aesthetic feelings stem from the harmonious interrelations inside the object and that the aesthetic measure is determined by the *order* relations in the aesthetic object.

While many of the examples given in this final section have attempted to relate entropy (as a measure of order) to perception in applications as diverse as the fine arts and in neuroarchitecture, in most cases the Rasch psychometric approach (Section 2.3.5) has yet to be deployed. As explained at the start of this section, unnecessarily large uncertainties may result, with substantial risks of incorrect decisions and potentially serious consequences in many fields of application. Luckily, as reviewed in this chapter, many of tools of modern measurement theory (Sections 2.2 and 2.3) are already accessible and applied to more basic signals and information, such as exemplified in Section 2.4.

When modern measurement theory will eventually be adopted, metrological standards, like words, in these new and topical applications should be able to mediate generally navigable conceptual ideals while providing unique local creative improvisations. Faithful communication of which quantities, units, and measures are meaningful is a basic requirement whenever inference and challenge are being tackled and however much order and complexity are involved.

References

Andersson, C., & Törnberg, P. (2018). Wickedness and the anatomy of complexity. *Futures*, *95*, 118–138. https://doi.org/10.1016/j.futures.2017.11.001

Armenia, K., Willems, R. M., & Frank, S. L. (2017). Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience and Biobehavioral Reviews*, *83*, 579–588.

Bashkansky, E., & Turetsky, V. (2016). Ability evaluation by binary tests: Problems, challenges & recent advances. *Journal of Physics: Conference Series*, *772*, 012012. https://doi.org/10.1088/1742-6596/772/1/012012

Bateson, G. 1979, *Mind and nature: A necessary unity*, New York: E. P. Dutton.

de Battisti, F., Nicolini, G., & Salini, S. (2010). The Rasch Model in customer satisfaction survey data. *Quality Technology & Quantitative Management*, *7*, 15.

Bense, M. (1969). *Einführung in die informationstheoretische Ästhetik. Grundlegung und Anwendung in der Texttheorie*. Reinbek bei Hamburg, Germany: Rowohlt Taschenbuch Verlag GmbH.

Bentley, J P. (2005). *Principles of measurement systems*. 4th edn, London: Pearson Education Limited.

Berglund, B., Rossi, G. B., Townsend, J., & Pendrill, L. R. (2012). (Ed.). *Theory and methods of measurements with persons*. Milton Park: Psychology Press, Taylor & Francis, ISBN 9781848729391.

Berglund, B., Nilsson, M. E., Sundling, C., Emardson, R., & Pendrill, L. R. (2014). Psychometric measurement and decision-making of accessibility in public transport for older persons with functional limitations, *Trafikverket rapport*, **TRV 2010/29710**, http://fudinfo.trafikverket.se/fudinfoexternwebb/pages/PublikationVisa.aspx?PublikationId=2020

Berlyne, D. E., Ogilvie, J. C., & Parham, L. C. C. (1968). The dimensionality of visual complexity, interestingness, and pleasingness. *Canadian Journal of Psychology*, *22*, 376–387. https://doi.org/10.1037/h0082777, PubMed: 5724480.

Berlyne, D. E. (1971). *Aesthetics and psychobiology*. New York: Appleton Century-Crofts.

Bezruczko, N. (1990). The construction and validation of a Rasch preference scale for design complexity: an aspect of aesthetic judgment. *Doctoral dissertation*, The University of Chicago

Bezruczko, N., Manderscheid, E., & Schroeder, D. H. (2016). MRI of an artistic judgment aptitude construct derived from Eysenck's K factor. *Psychology & Neuroscience*, *9*, 293.

Birdsall, T. (1973). *The theory of signal detectability: ROC curves and their character*. Technical Report College of Engineering, University of Michigan, https://deepblue.lib.umich.edu/handle/2027.42/3618

Birkhoff, G. D. (1933). *Aesthetic measure*. Cambridge, MA, USA: Harvard University Press.

Boon, J. P. (2011). *Artistic forms and complexity*. Nonlinear Dynamics, Psychology, and Life Sciences.

Born, M. (1972). *Atomic physics*. 8th ed. London, Glasgow: Blackie & Son Ltd, 216.89027.6, ISBN-13: 978–0486659848, ISBN-10: 0486659844.

Brillouin, L. (1962). Science and information theory. In *Physics today* Vol. 15, 2nd ed., Academic Press, https://doi.org/10.1063/1.3057866

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. 2nd ed., New York: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261–304. https://doi.org/10.1177/0049124104268644

Carnot, L. (1803) *Principes fondamentaux de l'équilibre et du mouvement*; de l'imprimerie de Crapelet.

Cedeño-Laurent, J., Williams, A., MacNaughton, P., Cao, X., Eitland, E., Spengler, J., & Allen, J. (2018). Building evidence for health: Green buildings, current science, and future challenges. *Annual Review of Public Health*, *39*, 291–308. https://doi.org/10.1146/annurev-publhealth-031816-044420, https://doi.org/10.1146/annurev-publhealth-031816-044420

CGPM 2018a, *SI Brochure: The International System of Units (SI)*, https://www.bipm.org/en/publications/si-brochure/

CGPM 2018b, “On the revision of the international system of units (SI)”, Draft Resolution A – 26th meeting of the CGPM (13–16 November 2018), https://www.bipm.org/utils/en/pdf/CGPM/Draft-Resolution-A-EN.pdf

Chaitin, G. J. (1969). On the length of programs for computing finite binary sequences. *Journal of the Association for Computing Machinery*, *13*, 547.

Chen, L., Liang, X., & Li, T. C. (2015). Collaborative performance research on multi-level hospital management based on synergy entropy-HoQ. *Entropy*, *17*, 2409–2431. https://doi.org/10.3390/e17042409

Church, R. L., & Marston, J. R. (2003). Measuring accessibility for people with a disability. *Geographical Analysis*, *35*, 83–96. https://doi.org/10.1353/geo.2002.0029

Cipriani, D., Fox, C., Khuder, S., & Boudreau, N. (2005). Comparing Rasch analyses probability estimates to sensitivity, specificity and likelihood ratios when examining the utility of medical diagnostic tests. *Journal of Applied Measurement*, *6*, 180–201.

Cohen, S., Kessler, R. C., & Gordon, L. U. (1995). Strategies for measuring stress in studies of psychiatric and physical disorders. In *Measuring stress: A guide for health and social scientists* (pp. 3–26). Oxford; New York, N.Y: Oxford University Press.

Cohen-Tannoudji, G. (2009). Universal constants, standard models and fundamental metrology. *European Physical Journal: Special Topics*, *172*, 5–24. https://doi.org/10.1140/epjst/e2009-01038-2

Commons, M. L., Gane-McCalla, R., Barker, C. D., & Li, E. Y. (2014). The model of hierarchical complexity as a measurement system. *Behavioral Development Bulletin*, *19*, 9–14.

Cordier, P., France, M. M., Pailhous, J., & Bolon, P. (1994). Entropy as a global variable of the learning process. *Human Movement Science*, *13*, 745–763.

de Courtenay, N. (2015). The double interpretation of the equations of physics and the quest for common meanings. In O. Schlaudt & L. Huber (Eds.), *Standardization in measurement* (pp. 53–68). London: Pickering & Chatto Publishers.

Dawson-Tunik, T. L., Commons, M., Wilson, M., & Fischer, K. W. (2005). The shape of development. *European Journal of Developmental Psychology*, *2*, 163–195. https://doi.org/10.1080/17405620544000011

Dirac, P. A. M. (1992). *The principles of quantum mechanics*. 4th ed. The International Series of Monographs on Physics. J. Birman, et al. (Ed.). Oxford: Clarendon Press.

Dybkaer, R. (2010). ISO terminological analysis of the VIM3 concepts ‘quantity’ and ‘kind-of-quantity’. *Metrologia*, *47*, 127–137. https://doi.org/10.1088/0026-1394/47/3/003

Dzhafarov, E. N. (2012). Mathematical foundations of universal Fechnerian scaling. In Berglund, et al. (Ed.). *Theory and methods of measurements with persons*.

Emardson, R., & Jarlemark, P. (2005). Uncertainty evaluation in multivariate analysis – a test case study. *Journal of Mathematical Modelling and Algorithms*, *4*, 289–305. http://dx.doi.org/10.1007/s10852-005-9005-2

Emerson, W. H. (2008). On quantity calculus and units of measurement. *Metrologia*, *45*, 134–138.

EU AI Act. (2023). https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Fava, G. A., Sonino, N., Lucente, M., & Guidi, J. (2022). Allostatic load in clinical practice. *Clinical Psychological Science*, *11*, 345–356. https://doi.org/10.1177/21677026221121601

Fernandez, D., & Wilkins, A. J. (2008). Uncomfortable images in art and nature. *Perception*, *37*, 1098–13.

Finkelstein, L. (1975). Fundamental concepts of measurement: Definition and scales. *Measurement and Control*, *8*, 105–111. https://doi.org/10.1177/002029407500800305G

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374. https://www.sciencedirect.com/science/article/pii/0001691873900036

Fisher, W. P., Jr, & Burton, E. (2010). Embedding measurement within existing computerized data systems: Scaling clinical laboratory and medical records heart failure data to predict ICU admission. *Journal of Applied Measurement*, *11*, 271–287.

Flater, D. (2018). Architecture for software-assisted quantity calculus. *Computer Standards and Interfaces*, *56*, https://doi.org/10.1016/j.csi.2017.10.002

Flater, D. (2023). Dealing with counts and other quantal quantities in quantity calculus. *Measurement*, *206*, https://doi.org/10.1016/j.measurement.2022.112226

Fleischmann, R. (1960). Einheiteninvariante Gröβengleichungen, Dimension. *Der Mathematische Und Naturwissenschaftliche Unterricht*, *12*, 386–399.

Forsythe, A., Nadal, M., Sheehy, N., & Cela-Conde, C. J. (2011). Predicting beauty: Fractal dimension and visual complexity in art. *British Journal of Psychology*, *102*, 49–70. https://doi.org/10.1348/000712610X498958, PubMed: 21241285.

Goel, A., & Chandrasekaran, B. (1992). Case-based design: A task analysis. In C. Tong & D. Sriram (Eds.). *Artificial intelligence approaches to engineering design*, *Vol. II: Innovative design* (pp. 165–184). San Diego: Academic Press.

McGrane, J. (2015). Stevens' forgotten crossroads: The divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Frontiers in Psychology*, *6*, 431–438. https://doi.org/10.3389/fpsyg.2015.00431

Granger, G. G. (1988). *Essai d'une philosophie du style* (pp. 71–105). Paris: Odile Jacob.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Huntington, NY: Krieger.

Green, K. E., & Smith, R. M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics*, *12*, 369. https://doi.org/10.2307/1165055

Guidi, J., Lucente, M., Sonino, N., & Fava, G. A. (2021). Allostatic load and its impact on health: A systematic review. *Psychotherapy and Psychosomatics*, *90*, 11–27. https://doi.org/10.1159/000510696

Guilford, J. P. (1936). *Psychometric methods*. New York: McGraw-Hill, Inc.

Hadžibajramović, E., Ahlborg, G., Håkansson, C., Lundgren-Nilsson, Å., & Grimby-Ekman, A. (2015). Affective stress responses during leisure time: Validity evaluation of a modified version of the stress-energy questionnaire. *Scand Journal Public Health*, 43(8), 825–832. https://doi.org/1403494815601552

Hartley, R. V. L. (1928). Transmission of Information. *Bell System Technical Journal*, *7*, 535–563.

van der Helm P A, (2000). Simplicity versus likelihood in visual perception: From surprisals to precisals. *Psychological Bulletin*, *126*, 770–800. https://doi.org/10.1037110033-2909.126.5.770

Hermans, M. C. E., Hoeijmakers, J. G. J., Faber, C. G., & Merkies, I. S. J. (2015). Reconstructing the rasch-built myotonic dystrophy type 1 activity and participation scale. *PLOS ONE*, *10*, e0139944. https://doi.org/10.1371/journal.pone.0139944

Hughes, L. F., Perkins, K., Wright, B. D., & Westrick, H. (2003). Using a Rasch scale to characterize the clinical features of patients with a clinical diagnosis of uncertain, probable, or possible Alzheimer disease at intake. *Journal of Alzheimer's Disease*, *5*, 367–373.

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological Bulletin*, *140*, 339–373. https://doi.org/10.1037/a0034221

Iosa, M., Bini, F., Marinozzi, F., Antonucci, G., Pascucci, S., Baghini, G., Guarino, V., Paolucci, S., Morone, G., & Tieri, G. (2022). Inside the Michelangelo effect: The role of art and aesthetic attractiveness on perceived fatigue and hand kinematics in virtual painting. *PsyCh Journal*, *11*, 748–54. https://doi.org/10.1002/pchj.606

ISO [704]:2022 *Terminology work – Principles and methods*, International Standardisation OrganisationISO 1087:2019 *Terminology work and terminology science: Vocabulary*

ISO/IEC 11179-31:2023 *Information technology – Metadata registries (MDR) – Part 31: Metamodel for data specification registration*, International Standardisation Organisation

ISO/IEC 21838-2:2021, *Information technology – Top-level ontologies (TLO) – Part 2: Basic Formal Ontology (BFO)*, https://www.iso.org/standard/74572.html

ISO/IEC 22989:2022 Information technology – Artificial intelligence – Artificial intelligence concepts and terminology

Iverson, G., & Luce, R. (1998). The representational measurement approach to psychophysical and judgmental problems. In M. H. Birnbaum (Ed.), *Measurement, judgment, and decision making*, (pp. 1–80). Academic Press.

Johnson, A. J., Dudley, W. N., Wideman, L., & Schulz, M. (2019). Physiological risk profiles and allostatic load: Using latent profile analysis to examine socioeconomic differences in physiological patterns of risk. *European Journal of Environment and Public Health*, *3*, em0029. https://doi.org/10.29333/ejeph/5870

Kellen, D., Davis-Stober, C. P., Dunn, J., & Kalish, M. (2021). The problem of coordination and the pursuit of structural constraints in psychology. *Perspectives on Psychological Science*, *16*, 767–778.

Khurana, A., & Bhatnagar, V. (2022). Investigating entropy for extractive document summarization. *Expert Systems with Applications*, *187*, 115820. https://doi.org/10.1016/j.eswa.2021.115820

Kim, D., Son, S. W., & Jeong, H. (2014). Large-scale quantitative analysis of painting arts. *Scientific Reports*, *4*, 7370. https://doi.org/10.1038/srep07370

Klir, G. J., & Folger, T. A. (1988). *Fuzzy sets, uncertainty and information*. New Jersey: Prentice Hall, ISBN 0-13-345984-5.

Knox, H. (1914). A scale, based on the work at Ellis Island, for estimating mental defect. *Journal of the American Medical Association*, *LXII*(10), 741–747.

Kolmogorov, A. N. (1965). Three approaches to the definition of the concept “amount of information”. *Problemy Peredachi Lnforrnatsii [Problems of Transmission of Information]*, *I*(1), 3–11. [in Russian].

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Vol. I: Additive and polynomial representations*. New York: Academic Press.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86. https://doi.org/10.1214/aoms/1177729694

Larsen, T. M., Endo, B. H., Yee, A. T., Do, T., & Lo, S. M. (2022). Probing internal assumptions of the revised bloom’s taxonomy. *CBE Life Sciences Education*, *21*, https://doi.org/10.1187/cbe.20-08-0170-CORRECTION

Linacre, J. M. (1994). Evaluating a screening test. *Rasch Measurement Transactions*, *7*, 317–318.

Linacre, J. M. (2006). Bernoulli trials, Fisher information, Shannon information and rasch. *Rasch Measurement Transactions*, *20*(3), 1062–1063. https://www.rasch.org/rmt/rmt203a.htm

Linacre, J. M., & Wright, B. (1989). The ‘length’ of a logit. *Rasch Measurement Transactions*, *3*, 54–55.

Luce, R. D. (1959). Individual choice behavior: A theoretical analysis. New York: Wiley.

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement*. New York: Springer.

Marin, M. M., & Leder, H. (2013). Examining complexity across domains: Relating subjective and objective measures of affective environmental scenes, paintings and music. *PLoS ONE*, *8*, e72412. https://doi.org/10.1371/journal.pone.0072412

Maroney, O. (2009). Information processing and thermodynamic entropy. In E. Zalta (Ed.). *Stanford Encyclopedia of Philosophy*. (Fall 2009 Edition). https://plato.stanford.edu/archives/fall2009/entries/information-entropy/

Maršik, F., & Mejsnar, J. (1994). The balance of entropy underlying muscle performance. *Journal of Non-Equilibrium Thermodynamics*, *19*, 197. https://doi.org/10.1515/jnet.1994.19.3.197

Massof, R. W. (2014). Are subscales compatible with univariate measures. In *International Objective Measurement Workshop* Program, *Rasch Measurement Transactions*, *27*(4), 1446. Philadelphia, PA, USA. https://rasch.org/rmt/rmt274.pdf.

Massof, R. W., Bradley, C., & McCarthy, A. M. (2024). Constructing a continuous latent disease state variable from clinical signs and symptoms. In W. P. Fisher, Jr., L. R. Pendrill (Eds.), *Models, measurement and metrology extending the SI* (pp. 269–300). De Gruyter.

Maxwell, J. C. (1871). Remarks on the mathematical classification of physical quantities. *Proceedings of the London Mathematical Society*, *s1–3*, 224–233.

Melin, J. (2024). Is validity a straightforward concept to be used in measurements in the human and social sciences?, In W. P. Fisher, Jr., L. R. Pendrill (Eds.), *Models, measurement and metrology extending the SI* (pp. 157–190). De Gruyter.

Melin, J., Cano, S., Flöel, A., Göschel, L., & Pendrill, L. R. (2022). The role of entropy in construct specification equations (CSE) to improve the validity of memory tests: Extension to word lists. *Entropy*, *24*, 934. https://doi.org/10.3390/e24070934

Melin, J., Cano, S. J., Flöel, A., Göschel, L., & Pendrill, L. R. (2022a). Metrological advancements in cognitive measurement: A worked example with the NeuroMET memory metric providing more reliability and efficiency. *Measurement: Sensors*, 100658. https://doi.org/10.1016/j.measen.2022.100658

Melin, J., Cano, S. J., Gillman, A., Marquis, S., Flöel, A., Göschel, L., & Pendrill, L. R. (2023). Traceability and comparability through crosswalks with the NeuroMET Memory Metric. *Scientific Reports. Nature Publishing Group*, *13*(1), 1–12. https://doi.org/10.1038/s41598-023-32208-0

Melin, J., Kettunen, P., Wallin, A., & Pendrill, L. R. (2022). Entropy-based explanations of serial position and learning effects in ordinal responses to word list tests. *IMEKO Conference*, Porto (PT).

Melin, J., Cano, S. J., Göschel, L., Fillmer, A., Lehmann, S., Hirtz, C., Flöel, A., & Pendrill, L. R. (2021). Construct specification equations: 'Recipes' for certified reference materials in cognitive measurement. *Measurement: Sensors*, *18*, 100290. https://doi.org/10.1016/j.measen.2021.100290, https://www.sciencedirect.com/science/article/pii/S2665917421002531

Melin, J., & Pendrill, L. R. (2023). The role of construct specification equations and entropy in the measurement of memory". *Person-Centered Outcome Metrology: Principles and Applications for High Stakes Decision Making*. In W. P. Fisher Jr & S. J. Cano (Eds.). *Springer series in measurement science and technology* (pp. 269–309). Available from https://doi.org/10.1007/978-3-031-07465-3_10

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol Rev*, *63*, 81–97.

Moles, A. (1968). *Information theory and esthetic perception*. Urbana, IL, USA: University of Illinois Press.

Montgomery, D. C. (1996). *introduction to statistical quality control*. 3rd ed. New York: Wiley, ISBN 978-0-471-30353-4.

Murari, A., Peluso, E., Cianfrani, F., Gaudio, P., & Lungaroni, M. (2019). On the use of entropy to improve model selection criteria. *Entropy*, *21*, 394. https://doi.org/10.3390/e21040394

Nadal, M., Munar, E., Marty, G., & Cela-Conde, C. J. (2010). Visual complexity and beauty appreciation: Explaining the divergence of results. *Empirical Studies of the Arts*, *28*, 173–191. https://doi.org/10.2190/EM.28.2.d.17

Nake, F. (1974). *Ästhetik als Informationsverarbeitung Grundlagen und Anwendungen der Informatik im Bereich ästhetischer Produktion und Kritik*. Wien, Austria: Springer-Verlag.

Nelson, R. R. (2015). Physics envy: Get over it. *Issues in Science & Technology*, *2XI*, http://issues.org/31-3/physics-envy-get-over-it/

Nerbonne, J., Heering, W., & Kleiweg, P. (1999). Edit distance and dialect proximity. In D. Sankoff & J. Kruskal (Eds.). *Introduction to reissue edition, time warps, string edits and macromolecules: The theory and practice of sequence comparison*. CSLI.

Newton, I. (1769). *Universal arithmetick: or, A treatise of arithmetical composition and resolution*. (J. Ralphson, S. Cunn, tr.). London: W. Johnston. https://archive.org/details/universalarithm00wildgoog/page/n21/mode/2up

Orlitsky, A., Santhanam, N. P., Viswanathan, K., & Zhang, J. (2006). Limits results on pattern entropy. *IEEE Transactions on Information Theory*, *52*, 2954–2964. https://doi.org/10.1109/TIT.2006.876351

Pele, O., & Werman, M. (2010). The quadratic-chi histogram distance family. *Computer Vision–ECCV*, *2010*, 749–762. www.cs.huji.ac.il/~werman/Papers/ECCV2010.pdf

Pendrill, L. R. (2008). Operating 'cost' characteristics in sampling by variable and attribute. *Accreditation and Quality Assurance*, *13*, 619–631. http://dx.doi.org/10.1007/s00769-008-0438-y

Pendrill, L. R. (2014a). Using measurement uncertainty in decision-making & conformity assessment. *Metrologia*, *51*, S206. https://doi.org/10.1088/0026-1394/51/4/S206

Pendrill, L. R. (2014b). Man as a measurement instrument. NCSLI Measure: The Journal of Measurement Science, 9, 24–35.

Pendrill, L. R. (2019). *Quality assured measurement – unification across social and physical sciences*. Springer Series in Measurement Science and Technology, ISBN: 978-3-030-28695-8 (e-book), https://doi.org/10.1007/978-3-030-28695-8

Pendrill, L. R., Espinoza, A., Wadman, J., Nilsask, F., Wretborn, J., Ekelund, U., & Pahlm, U. (2021). Reducing search times and entropy in hospital emergency departments with real-time location systems. *IISE Transactions on Healthcare Systems Engineering*, https://www.tandfonline.com/doi/full/10.1080/24725579.2021.1881660

Pendrill, L. R., & Fisher, W. P., Jr. (2013). Quantifying human response: Linking metrological and psychometric characterisations of man as a measurement instrument. *Joint IMEKO TC1-TC7-TC13 Symposium, Measurement across physical and behavioural sciences*, 4–6 September 2013, Genova, Palazzo Ducale (IT), *Journal of Physics: Conference Series*, *459*, 012057. doi:10.1088/1742-6596/459/1/012057

Pendrill, L. R., & Petersson, N. (2016). Metrology of human-based and other qualitative measurements. *Measurement Science and Technology*, *27*, 094003. https://doi.org/10.1088/0957-0233/27/9/094003

Pendrill, L. R., Melin, J., Stavelin, A., & Nordin, G. (2023). Modernising receiver operating characteristic (ROC) curves. *Algorithms*, *16*, 253. https://doi.org/10.3390/a16050253

Peptenatu, D., Andronache, I., Ahammer, H., Taylor, R., Liritzis, I., Radulovic, M., Ciobanu, B., Burcea, M., Perc, M., Pham, T. D., Tomić, B. M., Cîrstea, C. I., Lemeni, A. N., Gruia, A. K., Grecu, A., Marin, M., & Jelinek, H. F. (2022). Kolmogorov compression complexity may differentiate different schools of Orthodox iconography. *Scientific Reports*, *12*(1), 0743. https://doi.org/10.1038/s41598-022-12826-w, PMID: 35750777; PMCID: PMC9232591.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*, 237–255. https://www.rasch.org/memo24.htm

Peyrard, M. (2019). A simple theoretical model of importance for summarization, arXiv:1801.08991v2 [cs.CL], 6 Aug 2019.

Peterson, W., Birdsall, T., & Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, *4*, 171–212. https://doi.org/10.1109/TIT.1954.1057460

Pretty, J., Peacock, J., Sellens, M., & Griffin, M. (2005). The mental and physical health outcomes of green exercise. *International Journal of Environmental Health Research*, *15*, 319–337. PubMed: 16416750

Prisma Labs (2023), https://prisma-ai.com/lensa

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut. (reprint 1980, Chicago: University of Chicago Press).

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, *4*, *Contributions to Biology and Problems of Medicine*, (pp. 321–333). University of California Press.

Rexstad, E. (2001). Back cover of T.M. Shenk. In A. B. Franklin (Ed.). *Modeling in natural resource management*. Washington, D.C: Island Press.

Rice, S., Pendrill, L. R., Petersson, N., Nordlinder, J., & Farbrot, A. (2018). Rationale and Design of a Novel Method to Assess the Usability of Body-Worn Absorbent Incontinence Care Products by Caregivers. *Journal of Wound Ostomy & Continence Nursing*, *45*, 456–464. https://doi.org/10.1097/WON.0000000000000462.

Rigau, J., Feixas, M., & Sbert, M. (2007). Conceptualizing Birkhoff's aesthetic measure using Shannon entropy and kolmogorov complexity. In D. W. Cunningham, G. Meyer, & L. Neumann (Eds.). *Computational aesthetics in graphics, visualization, and imaging* (pp. 101–112), Eurographics Association.

Rossi-Arnaud, C., Pieroni, L., & Baddeley, A. (2006). Symmetry and binding in visuo-spatial working memory. *Neuroscience*, *139*, 393–400.

Rossi, G. B., & Crenna, F. (2016). Toward a formal theory of the measuring system, IMEKO 2016 TC1-TC7-TC13. *Journal of Physics: Conference Series*, *772*, 012010. https://doi.org/10.1088/1742-6596/772/1/012010

Scha, R., & Bod, R. (1993). Computationele esthetical. *Informatie En Informatiebeleid*, *11*, 54–63. English translation in http://iaaa.nl/rs/campestE.html

Schneider, T. D., Stormo, G. D., Gold, L., & Ehrenfeuch, A. (1986). The information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, *188*, 415–431. www.lecb.ncifcrf.gov/~toms/paper/schneider1986

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: A new way to display consensus sequences. *Nucleic Acids Research*, *18*, 6097–6100. https://doi.org/10.1093/nar/18.20.6097

Schnore, M. M., & Partington, J. T. (1967). Immediate memory for visual patterns: Symmetry and amount of information. *Psychonomic Science*, *8*, 421–422.

Shannon, C. (1949). The mathematical theory of communication. In C. Shannon, W. Weaver, *The mathematical theory of communication* (pp. 29–127). University of Illinois Press.

Shemesh, A., Leisman, G., Bar, M., & Grobman, Y. J. (2021). A neurocognitive study of the emotional impact of geometrical criteria of architectural space. *Architectural Science Review*, *64*, 1–14. https://doi.org/10.1080/00038628.2021.1940827

Sigaki, H. Y. D., Perc, M., & Ribeiro, H. V. (2018). History of art paintings through the lens of entropy and complexity. *Proceedings of the National Academy of Sciences U S A*, *115*, E8585–94. https://doi.org/10.1073/pnas.1800083115, Epub 2018 Aug 27. PMID: 30150384; PMCID: PMC6140488.

Sommer, K. D., & Siebert, B. R. L. (2006). Systematic approach to the modelling of measurements for uncertainty evaluation. *Metrologia*, *43*, S200–10. https://doi.org/10.1088/0026.1394/43/4/S06

Sonin, A. A. (1997). *The Physical Basis of Dimensional Analysis*, First Edition published 1997. Versions of this material have been distributed in 2.25 Advanced Fluid Mechanics and other courses at MIT since 1992.

Stenner, A. J., & Fisher, W. P., Jr. (2013). Metrological traceability in the social sciences: A model from reading measurement. *Journal of Physics: Conference Series*, *459*, 012025. https://doi.org/10.1088/1742-6596/459/1/012025

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, Vol. 103, 677–680. New Series.

Sundling, C., Nilsson, M. E., Hellqvist, S., Pendrill, L. R., Emardson, R., & Berglund, B. (2015). Travel behaviour change in old age: The role of critical incidents in public transport. *European Journal of Ageing*, 1–9. https://doi.org/10.1007/s10433-015-0358-8, http://link.springer.com/article/10.1007/s10433-015-0358-8, open access.

Tanaka, H., Adachi, H., Kazui, H., Ikeda, M., Kudo, T., & Nakamura, S. (2019). Detecting dementia from face in human-agent interaction. In *Proceedings of the ICMI 2019*, pp. 1–6

Tukey, J. A. (1986). Data analysis and behavioral science. In L. V. Jones (Ed.). *The collected works of John A. Tukey, Volume III, Philosophy and principles of data analysis: 1949 – 1964*, (pp. 187–390). Chapman & Hall. https://archive.org/details/collectedworksof0004tuke/

Uher, J. (2019). Data generation methods across the empirical sciences: Differences in the study phenomena's accessibility and the processes of data encoding. *Quality and Quantity*, *53*, 221–246. https://doi.org/10.1007/s11135-018-0744-3

Weaver, W. (1949). Recent contributions to the mathematical theory of communication. In C. Shannon, W. Weaver, *The mathematical theory of communication* (pp. 1–28). University of Illinois Press.

Wells, R. B. (2011). *Weaver's Model of Communication and its Implications*, https://webpages.uidaho.edu/rwells/techdocs/Weavers%20Model%20of%20Communication%20and%20Its%20Implications.pdf

Weitzner, D. S., & Calamia, M. (2020). Serial position effects on list learning tasks in mild cognitive Impairment and Alzheimer's Disease. *Neuropsychology*, *34*, 467–478. https://doi.org/10.1037/neu0000620

Weinberger, E. D. (2003). A theory of pragmatic information and its application to the quasi-species model of biological evolution. *Biosystems*, *66*, 105–119. http://arxiv.org/abs/nlin.AO/0105030

Wikipedia. (2021). *Entropy in thermodynamics and information theory*. Retrived 31 Jan 2021 from: https://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory

Yonelinas, A. P., Aly, M., Wang, W.-C., & Koen, J. D. (2010). Recollection and Familiarity: Examining controversial assumptions and new directions. *Hippocampus*, *20*, 1178–1194.

Zhu, Y., Liang, X., Batsis, J. A., & Roth, R. M. (2021). Exploring deep transfer learning techniques for Alzheimer's dementia detection. *Frontiers in Computer Science*, *3*, 624683.

Zidek, J. V., & van Eeden C, (2003). Uncertainty, entropy, variance and the effect of partial information. *Lecture Notes-Monograph Series*, *42*, 155–167. Mathematical Statistics and Applications: Festschrift for Constance van Eeden https://projecteuclid.org/download/pdf_1/euclid.lnms/1215091936

Zipf, G. K. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology*. Cambridge: Addison-Wesley Press.

Zurek, W. H. (1989). Algorithmic randomness and physical entropy. *Physical Review*, *A 40*(8), 4731–4751.

Matt Barney and Feynman Barney

3 Transdisciplinary measurement through AI: hybrid metrology and psychometrics powered by large language models

Abstract: Imagine a world where interdisciplinary scientific collaboration is revolutionized through universally recognized, precise measurements. This chapter introduces a cutting-edge methodology that integrates principles from artificial intelligence, metrology, and psychometrics. By engineering prompts in large language models like GPT-4, we focus on these synthetic AI "raters" on the measurand or attribute of interest. Rigorous empirical evaluations in both computer science and psychology evaluate people and AI with engineering-worthy standards. This chapter shows empirical evidence with ethical persuasive language and comparing the performance of AI systems, using multifaceted probabilistic models of measurement that ensure linearity, precision, and bias remediation in ways that will directly support extensions of the SI units into new domains. The chapter also explores physical metrology, illustrating how our approach can streamline traditionally complex processes such as evaluating electrical resistance in materials. Beyond technicalities, our methodology liberates scientific creativity and enforces ethical rigor. The chapter first presents the methodology, then its diverse applications, and finally, the ethical dimensions. By enabling a shared language of high-quality measurements, we pave the way for groundbreaking, interdisciplinary collaboration.

> *Artificial Intelligence is a tool, not a threat. It's going to be our partner. If we misuse it, it will be a risk. If we use it right, it can be our partner.* – Masayoshi Son, Founder, SoftBank

Keywords: Transdisciplinary measurement, artificial intelligence (AI), hybrid metrology, psychometrics, large language models (LLMs), GPT-4, synthetic AI raters, multifaceted probabilistic models, precision, SI units extension, digital exhaust, unobtrusive measurement, metrological standards, bias remediation, interval quantities, construct mapping, ethical persuasion, machine learning, cross-disciplinary collaboration

Note: With editorial assistance from OpenAI's GPT-4, all errors are the fault of the human at matt@xlnc.co. Special thanks to Feynman Barney for inspiring the physical metrology hypotheses.

Matt Barney, TruMind.ai, Stratham, NH, USA
Feynman Barney, Lawrence-Berkeley National Laboratory, University of California-Berkeley, Berkeley, CA, USA

https://doi.org/10.1515/9783111036496-003

3.1 Introduction

What if scientists from different fields could transcend their disciplinary boundaries and work together to tackle the world's most complex and urgent problems? The existing SI units are the means for a wide variety of cross-, multi-, and trans-disciplinary collaborations and correspondences in measurement, but what if the breadth and depth of those relationships could be greatly expanded? A new potential convergence of knowledge and methods could drive unprecedented advances, but only if the various fields' toolsets can be synthesized in ways that are more elegant, powerful, and inexpensive than today. This is the enigmatic challenge at the heart of transdisciplinary measurement: can we create a system that transcends disciplinary boundaries and enables seamless collaboration, all while ensuring better reliability, accuracy, precision, traceability, and usability of the measurements involved? And those better measurements will lead to better and quality-assured products, services, and processes throughout society (Pendrill, 2019).

One new powerful ally to improve measurement across the sciences is machine learning, commonly referred to as artificial intelligence (AI). With its unparalleled ability to process vast amounts of data, recognize patterns, and generate insights, AI is revolutionizing the way we approach many types of measurement. But how can we harness this power in a way that respects the unique needs and perspectives of each scientific discipline so that the emergent whole is greater than the sum of the individual components?

This chapter presents one approach with early evidence that we've partially solved this captivating puzzle. It proposes a hybrid methodology that merges the latest advancements in large language models (LLMs) from computer science with the enduring principles and practices from metrology, along with rule-based, metrologically oriented psychometrics. This unique combination paves the way for innovative forms of transdisciplinary measurement.

While the chapter will give examples applied in only two disciplines – computer science and psychology – it hypothesizes applications in more traditional metrological domains as well. What makes the proposed hybrid methodology unique is its applicability across various types of qualitatively different, passively collected, and unobtrusively inferred forms of raw information sometimes called "digital exhaust" (e.g., vibration signals, potential field signals, remote sensing signals, text, audio, images, or video). In addition, the methodology leverages forms of AI called "large language models" (Kasneci et al., 2023; Zhao et al., 2023), which, despite their quite recent broad visibility, have informed reading comprehension measurement for decades (Bezirhan & von Davier, 2023; Stenner, 2023) and offer new potentials for estimating metrologically viable quantities in relation to the rigorous quality standards of multifaceted measurement models (Adams et al., 1997; Linacre, 1994).

The use of unobtrusively obtained data inferred from observed behaviors and decisions has become much easier in recent years as mobile devices, the internet, and blockchain have become increasingly ubiquitous. For instance, audio, video, text, and data from Internet of things (IoT) devices (like gyrometer readings) are being used to remotely measure cognitive and behavioral dimensions of persons suffering from dementia (David et al., 2023).

But this new approach is not merely intended to measure a wider variety of raw data types that might meet traditional metrological standards. It may also contribute to improvements in traditional physical metrology, where nonphysical sources of uncertainty may impact measurement precision, as with time, for example. With time, there are systematic distortions in time measurement caused by lab-specific differences between national metrology institutes. Even though the formal SI definition of the standard second is exactly 9,192,631,770 periods of the radiation corresponding to a hyperfine transition of cesium-133 in its ground state, different labs return different results empirically (Tal, 2014). Even though different clock designs result in different tradeoffs between desiderata like frequency accuracy and stability, variability still exists between metrology labs even when using the same methods, which is why the SI and the Metre Convention include an international program, the Mutual Recognition Arrangement (MRA) (Comite International des Poids et Mesures (CIPM), 1999), for each physical quantity, including time. One might expect that the magnitudes of institute-specific distortions affecting the quality of their measurands would be quantified and accounted for in metrological quality assurance protocols, but time measurement practice has a long tradition of omitting outliers (often ones repeatedly reported by the same metrology lab).

Omitting outliers would mean that the Echelle Atomique Libre (EAL) would be estimated as an average of clock indications weighted by frequency-stability, which threatens the stability of world time (Tal, 2014, p. 8). There are systematic and random individual (micro), process (meso), and group (macro) uncertainties that drive variation in the quality of the measurands they produce, as can be revealed in the MRA. But metrologists, and the managers of labs that run atomic clocks, are only human – as is recognized in the role of the operator in the Measurement System Analysis approach (JCGM, 2020; see section 2.3.3 in Pendrill 2023). Thus, even the most precise physical measurement systems might benefit from this hybrid, AI-powered approach that accounts for distortions introduced by human involvement in the measurement process that blends metrology with Industrial-Organizational Psychology and Psychometrics.

To put our approach in perspective, we'll go on a journey through a new lens of the philosophy, history, theory, and practice of measurement to suggest a new trajectory for all forms of scientific measurement, with concrete examples and early evidence for the efficacy of the approach.

3.2 A vision for measurement across the sciences

As we grapple with the complexity of transdisciplinary measurement, we must first acknowledge the necessity of employing theory-grounded, explainable, and rigorous instrumentation before introducing new technologies that make for improved, faster, or less resource-intensive processes. The vastness of the raw data available and the diversity of the disciplines that could be involved in creating new instrumentation require an approach that is both grounded in theory and transparent in its application.

Despite sharing the common goal of generating linear, accurate, precise, valid, reliable, and traceable measurements, specialists in the fields of metrology and psychometrics have diverged over time, both in their approaches and their language (Berglund et al. 2013). Metrologists refer to "measurands," as the properties or quantities they intend to measure, while psychometricians focus on latent "constructs" that are hypothesized attributes or traits inferred from raw data. The disciplinary divergence is partially from the fact that the social sciences study "intangibles," whereas the physical, biological, and chemical sciences study "tangibles" and these qualitative differences in scientific challenge have spawned different approaches to measurement.

However, at their core, the concepts of measurands and constructs are fundamentally similar, as both represent the real-world phenomena we seek to understand. In this sense, probabilistic models for measurement that focus on constructing interval quantities from ordinal observations provide methods that align well with traditional metrological approaches (Fisher, 2009, 2022; Fisher & Cano, 2023; Mari & Wilson, 2014; Mari et al., 2023; Pendrill, 2019; Pendrill & Fisher, 2015). These models are identified, which means they allow us to construct measurements that are not dependent on specific samples or items but are instead objectively generalizable across different contexts, a feature that echoes the universal applicability sought in both metrology and psychometrics.

3.2.1 AI Helping bridge metrology and psychometrics

Advanced AI methods are helping to bridge the gap between these two measurement fields. AI, particularly LLMs, offer ways to automate the creation of measurement instruments in any discipline, building on existing methods of automatic item generation (Attali, 2018; Bejar et al., 2003; Embretson, 1999; Haladyna & Gierl, 2012; Hornke & Habon, 1986; Kosh et al., 2019; Poinstingl, 2009; Sonnleitner, 2008) and enabling a more universal approach. LLMs can be pretrained to use calibrated prompts as a sort of empirical reasoning engine, focused on the measurand or latent trait, to collect and evaluate data that might rise to the level of being a measurement. Prompts are natural language text samples that prime the knowledge and reasoning capability of a pretrained AI to perform a task that historically only people could perform. In this way, AI can serve as a common language and method, uniting metrologists and psychometricians in shared pursuit of understanding the magnitudes of the world around us.

New approaches to LLMs process and generate human-like analyses of numbers, text, audio, images, and video that offer many new pathways to creating measurements that were previously not possible. Of particular interest are OpenAI's GPT3.5 Turbo, the fastest-adopted technology in human history, and its improved version GPT-4 (Bubeck et al., 2023).

The path to coherence in measurement might begin from analyses of existing data, an available instrument, or a theory of the construct to be quantified. In any case, an approach taking various forms that can be generally characterized as construct mapping (Daniel & Embretson, 2010; De Boeck & Wilson, 2004; Embretson, 2010; Fischer, 1973; Stenner et al., 2013; Stone et al., 1999; Wilson, 2004, 2023) offers a powerful strategy for both physical and social scientists. It emphasizes the importance of an a priori theory about how physical, chemical, biological, psychological, or social processes might manifest at every level of the instrument's desired utility or use cases. This approach directs the creator of a construct model and associated instrument to proactively engineer linear information with fit-for-purpose accuracy and precision across the full range of potential variation. A construct map might be devised from a theory of how variation in some area manifests itself, or from existing data that reveal how items included on an existing instrument are meaningfully ordered (Wright, 1994); or the content of an available set of items might be examined for evidence that a single conceptual dimension unfolds in a potentially linear way.

Even though this approach was originally created in the context of educational and psychological constructs, metrologists and psychometricians agree that modeling of this kind taps the kind of reasoning processes that characterize science in general (e.g., Fisher, 2010; Fisher & Stenner, 2013, 2016; Pendrill & Fisher, 2015; Pendrill, 2019; Mari et al., 2023). This commonality came about in part because Rasch worked in an environment infused with values informed by Maxwell's method of physical analogy and perspective on modeling (Fisher, 2023; Fisher & Stenner, 2013).

When using the construct mapping approach, the metrologist or psychometrician carefully considers what content, process, or evidence would represent the measurand or construct at every level, including the extremes, from an absolute zero at the lower end, all the way to the highest levels of practical utility at the upper end. By designing LLM prompts inspired by these kinds of mapping guidelines and the content described by available theory, measurement researchers can experimentally evaluate alternative ways to operationalize the construct and its quantification. Many-faceted models (Adams et al., 1997; Linacre, 1994) are particularly well suited to this task, because they include not just the samples of persons and items participating in manifesting the measurand or construct but also additional factors that might affect the estimation of comparable measurements, such as the severity or leniency bias of synthetic or human raters judging a magnitude, or variations across organizational contexts that might introduce otherwise uncontrolled effects.

The LLM approach to AI that uses generative, pretrained transformers allows a metrologist or psychometrician to "program" or "prime" the AI to act as if it were an assessment or metrological instrument. Instead of using a computer programming language (e.g., Python and C++), LLM allows the measurement specialist to program the AI using natural human language. Pretrained forms of AI already have content-focused knowledge and reasoning capability that needs to be further "primed" in relation to the mapped construct to act as a trustworthy collector of data sufficient to the task of interval quantification. While the AI's data may not always rise to the quality levels we require in metrology and psychometrics, in the prototype described at the end of the chapter, we'll see that they not only often do, but can achieve unprecedented levels of utility when properly primed.

Because AI isn't perfect, the metrologist or psychometrician has to hypothesize a variety of ways that might lead to the successful estimation of a measurement. This translates into employing a variety of LLMs with slightly different parameters, training datasets, or algorithms in conjunction with different arrays of prompts intended to quantify the magnitude of a measurand or construct. Once their results conform to the model's quality standards, the LLM can serve as a trustworthy system of synthetic raters, providing different "camera angles" to estimate scale locations and associated uncertainties. In contrast with much current LLM practice, however, it is not enough to stop with generating these synthetic ratings. We must also experimentally confirm that they meet or exceed the minimum standards of a good instrument and so provide the construct and predictive validity needed for practical applications.

This is where automated multifaceted models become invaluable. These models allow the measurement specialist to reject hypotheses about automatically generated prompts and ratings that don't fit the quality standards for objective measurement and also adjust for various synthetic (AI algorithm) rater biases, ensuring sufficient linearity, accuracy, precision, and freedom from unidentified sources of differential functioning across items, raters, persons, and groups (Andrich & Hagquist, 2012, 2015; Xie & Wilson, 2008).

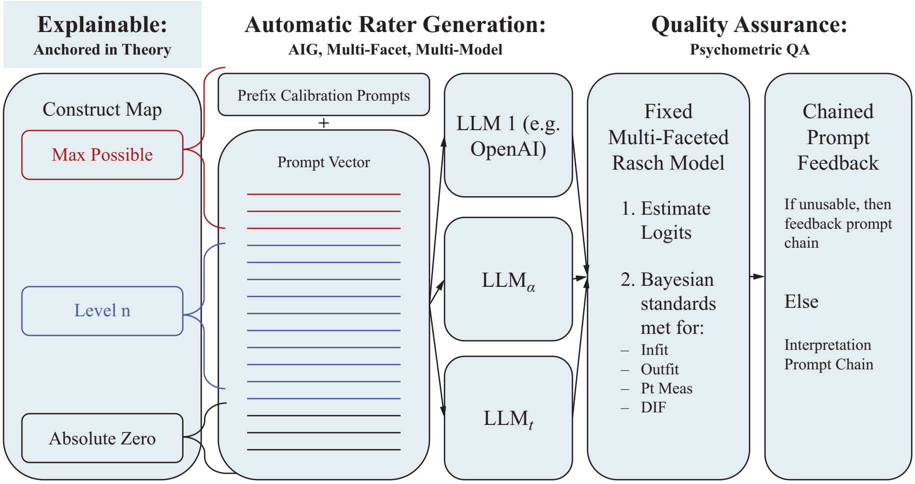


Figure 3.1: Explainable, unbiased measures with quality control and feedback.

Figure 3.1 shows an example of how the scientific conception of a construct would be mapped to ensure that the prompts used with an LLM are traceable to theory, that the prompt vectors are submitted to an array of LLMs of intentionally varied quality, cost, and response latency, and that an automated model evaluation identifies measurements that are distorted, nonlinear, imprecise, or otherwise not fit for purpose. In the very end, prompts can also be created to provide feedback helping the user to interpret the results and make improvements based on the measurements.

In summary, the vision for metrologically oriented psychometrics is a fusion of theory-grounded AI and transdisciplinary measurement approaches. It's a marriage of rigorous AI instrumentation, a priori theoretical grounding, and sophisticated measurement techniques. Together, they promise a future where scientific collaboration can thrive, grounded in the shared language of precise, reliable, and meaningful measurement.

3.2.2 Early evidence

The first author is an industrial-organizational psychologist who has been mentored by Dr. Robert Cialdini, a leading authority on ethical influence. He was honored to be acknowledged by the Society of Industrial-Organizational Psychology (American Psychological Association division 14) and the Association of Test Publishers in 2018 for work blending computer science and psychometric approaches. Yet, earlier attempts at creating an "Instant Persuasion Coach" with earlier deep learning transformer models ultimately failed.

The biggest challenge concerned the availability of the massive datasets required to fully train deep learning models. While large by psychological standards they were insufficient for traditional machine learning model requirements. For example, it was disappointing to find from a dataset of 4,500 CEO-investor interactions that the CEOs were almost universally poor at persuasion, and very few were even moderately good. Worse, 4,500 is a small sample by computer science standards. Consequently, the team failed to finish creating the "Instant Persuasion Assessment and Coaching" system we sought.

That changed in November of 2022. The introduction of ChatGPT by OpenAI brought about revolutionary possibilities not previously considered. LLMs, such as ChatGPT3.5Turbo, offered an entirely new paradigm that effectively addressed the struggles we had encountered in previous attempts at leveraging AI for measurement. Because they already had a sizable portion of the internet used to pretrain them, no longer are big datasets essential for every measurement situation. We quickly realized that by priming LLMs that already had reasoning and knowledge, but needed to be focused, we could achieve the results we desired without massive datasets, expense, or delays. Early anecdotal tests convinced us that it would not be a wasted effort to create an array of prompts, in much the same way as is traditionally done for computer-adaptive testing (Barney & Fisher, 2016). By grounding the prompts in Cialdini's model of ethical persuasion, together with both human and AI-generated samples of persuasion at multiple levels of ethics, and at multiple levels of Cialdini's first principle, reciprocity, it became clear that it was a much better paradigm.

For the study of ethical influence, we meticulously crafted 190 prompts, not merely using theory and evidence from my field of industrial-organizational psychology, nor only from Dr. Cialdini and his social psychological colleagues, but also from every discipline that had studied ethics. This included diverse fields, including computational linguistics, social psychology, philosophy, cultural anthropology, and communications, tapped for various ways to detect deceit, dishonesty, misdirection, and fake news (e.g., Boyd, Ashokkumar, Seraj & Pennebaker, 2022). Cialdini's second principle involves making sure that one persuades only things that are naturally present and are not contrived even if they are, strictly speaking, true, so that no one feels tricked or misguided. These prompts too were inspired across a wide array of disciplines including signal detection theory, Bayesian ideal observer theory, and the statistical properties of natural scenes (Geisler, 2008). For Cialdini's third and final dimension of ethical influence, wisdom, we considered interdisciplinary definitions of wisdom, as well as a wide array from different philosophical conceptions of wisdom in ancient and modern cultures (e.g., Staudinger & Gluck, 2011).

We integrated 190 prompts into a Google Sheet that combined them with five different settings called "temperatures" in GPT3.5 Turbo's Davinci algorithm to analyze a dozen human and AI-generated samples that exhibited varying levels of ethical reciprocity, lack of contrivance, and wisdom. Each level of influence sampled was processed using OpenAI's API for GPT-3, with the DaVinci algorithm. After collecting these observations, we estimated the parameters of a multifaceted measurement model (Linacre, 2022) to calibrate the instrument. Cronbach's alpha approached 1.00 and population separation and strata coefficients (Wright, 1996; Wright & Masters, 1982, pp. 92, 105–106) were between 32 and 44, suggesting over 30 distinguishable levels (ranges with centers three standard errors apart, corresponding to 99% level of statistical confidence). It worked with the first pass and at a relatively high level of measurement precision. Previous work with Cialdini using automatic item generation in computer adaptive tests had produced – after many years of refinement – a reliability of about 0.98.

Crucially, Cialdini's Influence Assessment (CIA) is a situational judgment test that takes 30–40 min to complete and is cognitively taxing, whereas this new LLM approach, when combined with an earlier technique called "inverted computer adaptive testing" can measure the same dimensions as the CIA with zero effort in six seconds of processing time. These promising results reaffirmed the conviction that this unobtrusive paradigm offers enormous potential. Looking ahead to the development of future LLMs capable of processing visual and auditory inputs, it is reasonable to anticipate even greater precision in measurement, particularly in high-context cultures where nonverbal cues play a significant role in estimating the degree to which persuasion is being used effectively.

This early prototype convinced us that a new paradigm in measurement science is emerging, one that is deeply rooted in longstanding values in the philosophy of science and ethical practices, and is capable of delivering valuable, accurate, and meaningful measurements – a capacity that scientists have been foreseeing for many years (e.g., Andrich, 1988; Bradley & Terry, 1952; Guttman, 1971; Luce, 1959; Loevinger, 1947, 1965; Luce & Tukey, 1964; Narens & Luce, 1986; Rasch, 1960; Thurstone, 1928; Wright, 1977, 1997; Fisher & Stenner, 2013) but which has not yet delivered the promised revolution (Cliff, 1992). It would seem that there is potential for efficiently and effectively extending the SI into previously excluded domains.

3.3 The role of transdisciplinary prompts in data collection

At the heart of this innovation in measurement is a new way to frame questions using prompts. What exactly is a prompt? For use in measurement, an AI prompt is a carefully constructed stimulus that evokes a response or output from a pretrained AI algorithm. The data produced by a prompt may rise to the level of informing the estimation of a measurement if it meets stringent standards. Analyzable output from a prompt that an LLM could process might be anything numerical, such as signals from an engineering, financial system, or a biological data collection device, or written text, audio, images, or video of a physical, chemical, biological, or human phenomenon. The goal in writing or "engineering" a prompt is to elicit an output that can be scored, scaled, and traced to the construct or property (the measurand) we are interested in measuring. This means the outputs must possess nonlinear and stochastic qualities that exhibit the kinds of spontaneously self-organized patterns proven in theory (Andersen, 1999; Andrich, 2010; Fischer, 1981) and practice as necessary and sufficient to the estimation of usefully unidimensional and linear measurements providing fit-for-purpose accuracy and precision. Although these patterns cannot be imposed by artificial contrivances, and though they are not automatically produced by just any collection of scored outputs, successful results are usually obtained by methodically organizing the observational frame of reference in terms of an appropriately structured construct theory.

We will explore this concept further through two empirical examples, one from computer science measurement and the other from psychometrics, before illustrating its potential application in physical metrology (section 3.4.1).

One chronic challenge in computer science has involved effective comparisons of the performance of various artificially intelligent systems. The best systems perform at super-human levels on everything from law to medical examinations, with well-designed psychometrics, but the same quality of measurement is not typically obtained when systems are evaluated in relation to a wide range of actual computer science variables (Bubeck et al., 2023). Many high-profile papers and AI leaderboards resort to counting percentages of correct responses, ignoring task difficulty, raw data distortions, nonlinearity, traceability, and other very basic metrological or psychometric concerns (e.g., HuggingFace 2023; Patil et al., 2023). One recent paper by Google that specifically focused on evaluation of AI not only did not include metrological or psychometric approaches to defining interval quantities (Mari & Wilson 2014; Mari et al., 2023; Pendrill, 2019; Pendrill & Fisher, 2015) but also resorted to simple tallies and percentages in its attempt to suggest improvements (Gehrmann et al., 2022). Sadly, concerns about risks to AI safety do not apply state-of-the-art measurement in risk mitigation or propose how the rigorous methods of metrology or psychometrics could be put to productive use, as is evident in a recent paper (Shevlane et al., 2023).

3.3.1 Caring for our technologies as we do our children

One of the great shortcomings of the modern world is its tendency to apply different values in technical versus human contexts (Habermas, 1995; Haraway, 2022; Norris, 1999; Postman, 1992; Snow, 1964). Caring is assumed to involve subjective attributes applicable to matters of human interest but irrelevant to technical concerns with objective facts and data. This characterization persists despite the documented coproduction of science and society (Blok et al., 2020; Jasanoff, 2004; Bowker et al., 2014; Douglas, 1986), and the coevolution of the arts and sciences, with artists' and scientists' comparable levels of and kinds of creativity (Bullot et al., 2017). The assumption that technical effects exist naturally in a manner completely independent of human interests can lead those involved in their production to ignore or devalue some or all of their social and moral consequences or presume those consequences to be unavoidable or necessary. Even innovations as seemingly innocuous and beneficial as air conditioning or digging a well for a village that has a long tradition of walking miles for water may result in significant distortions of human relationships and realities. More complex failed and partially successful efforts at improving the human condition may be aptly considered monstrous (Haraway, 1992; Kristeva, 2012; Latour, 2012; Scott, 1998), including those associated with AI (Barney & Fisher, 2017; Dove & Fayard, 2020).

In contemplating, then, the measurement of risks associated with AI, a particular kind of uncaring inattention that often characterizes mainstream approaches to quantification, presents a notable risk of its own: the confusion of numeric counts for measured quantities (Bateson, 1978; Fisher, 2021; Wright, 1994). Although virtually everyone can understand that it is impossible to tell who has more rock or ability from simple counts of stones or correct answers devoid of additional information on the size of the stones or the difficulty of the questions, a large proportion of social and psychological measurement applications do not distinguish levels of complexity in either theory or practice (Dawson et al., 2006; Fisher, 2023; Rousseau, 1985). But, as noted by Star and Ruhleder (1996, p. 118), if we:

> design messaging systems blind to the discontinuous nature of the different levels of context, we end up with organizations which are split and confused, systems which are unused or circumvented, and a set of circumstances of our own creation which more deeply impress disparities on the organizational landscape.

It would seem imperative that these problems be avoided in the context of managing and mitigating the risks of AI. Quantitative methods play key roles in maintaining the systemic disempowerment characterizing today's educational, health care, governance, and other institutions (Merry, 2016). Shifting the paradigm toward capacities for empowerment requires attending closely to the origins and destinations of the information communicated within and across organizational levels of complexity.

To show the utility of our approach in both computer science and social psychology, we prototyped an approach based on metrologically oriented psychometrics. We included two different LLMs, OpenAI's GPT-3.5 and GPT-4, along with different temperature settings that affect the stochasticity of the model. This exploratory analysis was intended to evaluate prompts designed to measure applications of Cialdini's principle of reciprocity, a well-known phenomenon established in social psychology as an essential component of relationship cultivation (Cialdini, 2009, 2018, 2021). To address the limits of the available computer science, we examined the degree to which simpler, cheaper LLMs were sufficiently useful for measurement in comparison with the more expensive, larger LLMs, measuring both on the same "ruler" as the experimental prompts created with the construct mapping approach.

In the first test, we queried all samples with every possible combination of GPT-3.5 with different temperature settings (0, 0.5, 1) and GPT-4 with one temperature setting (0, the most predictable), submitting 28 different prompts mapped to Cialdini's "activators" and "amplifiers" of persuasion for 14 different samples. To make sure the samples covered the full range of interest, some were authored by an expert in Cialdini methods and others were generated by GPT3.5. The Wright map in Figure 3.2 shows that, though the different AI algorithms and temperatures were largely similar, there was wide variability present in prompt (item) and sample locations. Because every facet element was used on every sample, they're all in the same frame of reference. While it's not entirely clear why the temperatures seemed to make no difference, it was useful to know so that we could move on to other possible sources of uncertainty in LLMs. LLMs themselves are statistical and dustbowl empirical, so it is somewhat surprising that a warmer temperature that should be more stochastic didn't have a different effect than a cold temperature that is relatively more deterministic. It is possible that our sample was too small to see much of a difference, or that these differences are just really small to begin with.

Usefully, Table 3.1 shows more details on whether the LLM's data exhibits structures approximately satisfying the constraints imposed by the measurement model. Information-weighted and outlier-sensitive mean-square fit statistics are ideally 1.00 (Linacre, 2003; Linacre & Wright, 1994), but, as Rasch stressed, "all statistical models are wrong" (Rasch, 1973/2011) and "a model is not meant to be true" (Rasch, 1960, pp. 37–38). Data never fit measurement models perfectly. The point of making models is to obtain tools and information useful with respect to a particular application, such that inevitably present uncertainty and imprecision does not rise to a level compromising the tool's status as fit for purpose.

```
+-------------------------------------------------------------------------------------+
|Measr|+Algorithm                                 |+Sample      |+Source|+Item  |AGREE|
|-----+-------------------------------------------+-------------+-------+-------+-----|
|   5 +                                           +             +       +       +(10) |
|     |                                           |             |       |       |     |
|     |                                           |             |       |       |     |
|     |                                           |             |       |       |     |
|     |                                           | GPT4 HRLE   |       |       |     |
|   4 +                                           +             +       +       +     |
|     |                                           |             |       |       |     |
|     |                                           |             |       |       |     |
|     |                                           | GPT4 HRHE   |       |       |     |
|     |                                           |             |       |       |     |
|   3 +                                           + Turbo Frank +       +       + --- |
|     |                                           |             |       |       |     |
|     |                                           |             |       |       |     |
|     |                                           | GPT4 MRHE   |       |       |     |
|     |                                           |             |       |       |     |
|   2 +                                           +             +       +       +  8  |
|     |                                           | Frank's Lady|       | *     |     |
|     |                                           |             |       |       |     |
|     |                                           | Daycare     |       |       |     |
|     |                                           |             |       | *     |     |
|   1 +                                           +             +       + **    + --- |
|     |                                           |             |       | ***   |     |
|     |                                           | Merck Food  | Human |       |     |
|     |                                           | CEO         |       | *     |     |
|     | T0 g3.5 turbo T0.5 3.5 turbo T1 3.5 turbo |             |       | ***** |  4  |
*   0 *                                           *             *       *       *     *
|     |                                           | Spence      |       | *     |     |
|     | GPT-4                                     |             |       | **    |     |
|     |                                           | GPT4 MRLE   | GPT4  | *     | --- |
|     |                                           |             |       | *     |     |
|  -1 +                                           +             +       + *     +     |
|     |                                           |             |       |       |     |
|     |                                           |             |       |       |     |
|     |                                           |             |       | *     |     |
|     |                                           |             |       | *     |  2  |
|  -2 +                                           +             +       + *     +     |
|     |                                           |             |       |       |     |
|     |                                           |             |       |       |     |
|     |                                           |             |       |       |     |
|     |                                           |             |       |       |     |
|  -3 +                                           +             +       +       +     |
|     |                                           |             |       |       |     |
|     |                                           | GPT4 LRLE   |       |       | --- |
|     |                                           |             |       |       |     |
|     |                                           |             |       |       |     |
|  -4 +                                           +             +       +       + (1) |
|-----+-------------------------------------------+-------------+-------+-------+-----|
|Measr|+Algorithm                                 |+Sample      |+Source| * = 1 |AGREE|
+-------------------------------------------------------------------------------------+
```

Figure 3.2: Wright map showing AI algorithms, prompts, samples, and sources on the same linear, unidimensional measure for Cialdini's measure of reciprocity.

Useful and meaningful results may then often be obtained even when statistical indicators flag some observations as significantly different departures from modeled expectations. For instance, it may happen that experiments provisionally omitting observations associated with mean square fit statistics exceeding a desired limit result in no change outside the bounds of the estimated uncertainty to either the estimated ability measurements or the item difficulties. Given the lack of a difference that makes a difference, the anomalous observations should be retained and fed back end users, since the unexpected responses may convey information valuable in the formulation of interventions, treatments, or decisions (Allen & Pak, 2023; Bohlig et al., 1998; Fisher et al., 2021; Massof & Bradley, 2023; Wilson & Gochyyev, 2020).

Table 3.1: Comparing large language models and temperatures.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Algorithm
1358	242	5.61	5.28	-0.47	0.09	1.19	2.0	1.27	2.7	0.7	0.72	0.81	54.5	41.1	GPT-4
1571	242	6.49	6.43	0.13	0.1	1.03	0.3	1.06	0.5	0.95	0.83	0.83	57.1	43.0	T1 3.5 turbo
1603	242	6.62	6.49	0.16	0.1	0.94	-0.6	0.91	-0.8	1.07	0.84	0.82	62.6	43.4	T0.5 3.5 turbo
1627	242	6.72	6.54	0.18	0.1	0.74	-3.1	0.73	-2.9	1.27	0.87	0.82	68.2	43.4	T0 g3.5 turbo
1539.8	242.0	6.36	6.19	.00	.10	.97	-.3	.99	-.1		.82				Mean
106.8	.0	.44	.53	.27	.00	.16	1.9	.20	2.1		.06				S.D. (Population)
123.3	.0	.51	.61	.31	.00	.19	2.2	.23	2.4		.07				S.D. (Sample)

Model, Populn: RMSE .10 Adj (True) S.D. .25 Separation 2.66 Strata 3.88 Reliability (not inter-rater) .88
Model, Sample: RMSE .10 Adj (True) S.D. .30 Separation 3.13 Strata 4.50 Reliability (not inter-rater) .91
Model, Fixed (all same) chi-squared: 33.6 d.f.: 3 significance (probability): .00
Model, Random (normal) chi-squared: 2.8 d.f.: 2 significance (probability): .25
Inter-Rater agreement opportunities: 1386 Exact agreements: 841 = 60.7% Expected: 592.0 = 42.7%

When considering the big picture, as shown in the Wright map in Figure 3.2, while the LLMs are relatively more like each other than prompts are, there are significant differences between them. This can be seen by the separation and strata values (population estimates of 2.66 and 3.88 respectively) suggesting that the most advanced LLM, GPT-4, is more lenient than all the various temperatures of GPT-3 that were sampled.

In this way, the MFRM allows us to compare not only the measurements relevant to computer science about the efficacy of different LLMs and settings to create good instruments (and evaluate LLM bias) but also the psychological variables represented by Cialdini's model of reciprocity and the cultivation of relationships in an ethical influence process. Table 3.2 shows measures of each sample on the same ruler. The fit indicators show noisier measures with more uncertainty for the lowest performing sample of reciprocity that was proactively written to be low for this test (#60, −3.44 log-odds units), and near-perfect information for sample 21 that was written using GPT-4 to be moderately persuasive, and ethical. Notice that the attempt at measurement achieved very good separation (12.68) and strata (17.24) and reliability (0.99).

Finally, we also were able to create a good set of prompts and evaluate and remove those that do not live up to the standards of interval-quality measurement to create the final instrument in Table 3.3. In this way, we could combine results from various computer science settings to evaluate alternative approaches, psychological samples, and create multiple instruments addressing different purposes. The end product is not perfect and is always accompanied with estimated uncertainty and data quality statistics, but careful and rigorous testing shows the measurements to be linear, accurate, precise, traceable to theory, and useful within the tolerance limits of the relevant applications.

3.3.2 Hypothesized uses for other latent constructs

To further show the value of our interdisciplinary approach, we will show hypothesized uses with example prompts that could be tested in the same way.

A key advantage of using an LLM is that it can be used to analyze data more quickly and consistently than is possible for a human. A metrological approach to leveraging an LLM is also less likely to make mistakes, as it is not affected by fatigue or boredom. Additionally, the LLM can be used to analyze data from a variety of sources, which can improve the accuracy of the measurements, when they conform to measurement quality standards.

Second, consider the task of measuring the personality trait of conscientiousness from a sample of digital text exhaust from an employee applying for a job. For over 100 years, conscientiousness has been shown in the literature as one of the most important predictors of job performance in all jobs (including but not limited to metrologists), and predicts organizational citizenship behavior, counterproductive work

Table 3.2: Measuring persuasive samples for reciprocity.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Sample
152	88	1.73	1.5	-3.44	0.2	1.41	2.3	1.48	2.6	0.48	0.24	0.51	GPT4 LRLE
225	88	2.56	3.24	-0.59	0.16	0.62	-2.7	0.65	-2.5	1.36	0.39	0.53	GPT4 MRLE
448	88	5.09	3.7	-0.22	0.14	0.77	-1.7	0.82	-1.3	1.16	0.66	0.6	Spence
562	88	6.39	4.81	0.47	0.14	1.14	0.9	1.18	1.2	0.66	0.78	0.61	CEO
554	88	6.3	5.03	0.59	0.14	0.83	-1.2	0.83	-1.2	1.24	0.69	0.61	Merck Food
668	88	7.59	6.68	1.45	0.15	0.68	-2.4	0.67	-2.4	1.42	0.75	0.6	Daycare
712	88	8.09	7.49	1.89	0.16	1.1	0.6	1.09	0.5	0.97	0.56	0.58	Frank's Lady
624	88	7.09	8.21	2.35	0.14	1.06	0.4	1.05	0.3	0.95	0.67	0.61	GPT4 MRHE
686	88	7.8	8.93	2.98	0.16	1.11	0.8	1.14	0.8	0.84	0.45	0.6	Turbo Frank
720	88	8.18	9.22	3.34	0.16	1.39	2.2	1.32	1.7	0.68	0.43	0.58	GPT4 HRHE
808	88	9.18	9.68	4.29	0.2	0.83	-0.9	0.71	-1.2	1.13	0.56	0.5	GPT4 HRLE
559.9	88.0	6.36	6.22	1.19	.16	.99	-.1	.99	-.1		.56		Mean
198.4	.0	2.25	2.61	2.06	.02	.26	1.7	.26	1.7		.16		S.D. (Population)
208.1	.0	2.36	2.74	2.16	.02	.27	1.8	.28	1.7		.17		S.D. (Sample)

Model, Populn: RMSE .16 Adj (True) S.D. 2.06 Separation 12.68 Strata 17.24 Reliability .99

Model, Sample: RMSE .16 Adj (True) S.D. 2.16 Separation 13.30 Strata 18.07 Reliability .99

Model, Fixed (all same) chi-squared: 1433.2 d.f.: 10 significance (probability): .00

Model, Random (normal) chi-squared: 9.9 d.f.: 9 significance (probability): .36

Table 3.3: Prompt measurement.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Prompt
173	44	3.93	3.0	-1.99	0.22	1.2	0.9	1.26	1.1	0.68	0.67	0.72	ACT:Gift Expert Conservative ARG
194	44	4.41	3.31	-1.72	0.21	1.99	3.7	2.06	3.8	-0.3	0.68	0.74	General ARG
175	44	3.98	3.48	-1.58	0.21	0.69	-1.6	0.79	-0.9	1.23	0.7	0.74	ACT:Concession#2 ARG
212	44	4.82	4.24	-1.05	0.21	0.68	-1.6	0.82	-0.8	1.22	0.73	0.76	AMP:Personalized#1 SS
234	44	5.32	4.83	-0.71	0.21	1.08	0.4	1.33	1.5	0.59	0.6	0.78	ACT:Concession In the moment#8 SS
251	44	5.7	5.07	-0.58	0.21	0.51	-2.8	0.54	-2.6	1.45	0.86	0.78	AMP:Personalized#3 SS
256	44	5.82	5.4	-0.4	0.21	1.24	1.1	1.13	0.6	0.97	0.83	0.79	General Severe SS
269	44	6.11	5.49	-0.36	0.21	0.72	-1.4	0.75	-1.2	1.14	0.78	0.79	AMP:Unexpected#1 SS
271	44	6.16	5.75	-0.22	0.21	1.43	1.9	1.34	1.5	0.58	0.73	0.79	ACT:Concession In the moment#7 SS
293	44	6.66	6.47	0.15	0.22	1.07	0.4	0.98	0.0	1.12	0.85	0.81	ACT:Gift#3 SS
293	44	6.66	6.47	0.15	0.22	0.74	-1.3	0.77	-1.0	1.33	0.87	0.81	ACT:Cooperation#1 SS
296	44	6.73	6.56	0.19	0.22	0.71	-1.4	0.77	-1.1	1.34	0.89	0.81	AMP:Significant#3 SS
303	44	6.89	6.66	0.24	0.22	1.17	0.8	1.22	1.0	0.59	0.74	0.81	ACT:Cooperation#2 Bard SS
291	44	6.61	6.75	0.29	0.22	1.0	0.0	1.0	0.0	1.03	0.79	0.81	AMP:Personalized#2 SS
308	44	7.0	6.84	0.34	0.22	0.81	-0.9	0.87	-0.5	1.2	0.88	0.81	ACT:Gift+Praise#1 SS
322	44	7.32	7.59	0.76	0.23	0.78	-1.0	0.75	-1.0	1.22	0.83	0.82	AMP:Unexpected#2 SS
325	44	7.39	7.59	0.76	0.23	1.0	0.0	1.07	0.3	0.99	0.84	0.82	AMP:Unexpected#3 SS
324	44	7.36	7.77	0.87	0.24	0.98	0.0	0.93	-0.2	1.18	0.83	0.82	ACT:Concession Compromise#4 SS
330	44	7.5	7.86	0.93	0.24	0.94	-0.1	0.88	-0.4	1.17	0.83	0.82	ACT:Concession#3 SS
331	44	7.52	7.86	0.93	0.24	0.91	-0.3	0.88	-0.4	1.22	0.89	0.82	ACT:Concession Accommodate#5 SS
344	44	7.82	8.3	1.23	0.25	1.18	0.7	1.21	0.8	0.87	0.78	0.82	ACT:Concession#1 SS
364	44	8.27	8.92	1.78	0.28	0.53	-2.0	0.51	-1.6	1.27	0.89	0.82	AMP:Significant#1 SS

280.0	44.0	6.36	6.19	.00	.22	.97	-.2	.99	-.1	.80	Mean
53.2	.0	1.21	1.64	.96	.02	.32	1.5	.33	1.4	.08	S.D. (Population)
54.4	.0	1.24	1.68	.98	.02	.33	1.5	.33	1.4	.08	S.D. (Sample)

Model, Populn: RMSE .22 Adj (True) S.D. .93 Separation 4.15 Strata 5.87 Reliability .95

Model, Sample: RMSE .22 Adj (True) S.D. .96 Separation 4.26 Strata 6.01 Reliability .95

Model, Fixed (all same) chi-squared: 387.8 d.f.: 21 significance (probability): .00

Model, Random (normal) chi-squared: 19.9 d.f.: 20 significance (probability): .46

behavior, job satisfaction, and likelihood of promotion (Wilmot & Ones, 2019). In this hypothesized situation, the prompt might be a specific question or statement designed to elicit a text response that reveals the person's level of conscientiousness, or from digital exhaust from a pre-recorded corpus of all the interviews the candidate had (e.g., downloaded from Zoom transcripts). For example, a prompt could be:

> *Act as an expert in personality psychology specializing in conscientiousness. Assess the following sample (text, imagery and/or audio) to evaluate the magnitude of every sub-facet including order, dutifulness, achievement striving, self-discipline, and deliberation. Rate their level of conscientiousness on a scale from 1 to 10, where:*
>
> 1 *Represents the lowest possible level of conscientiousness. The individual disregards details, exhibits chaotic behavior, lacks discipline, shirks responsibilities, and fails to deliver on commitments.*
>
> 2 *Shows minimal conscientiousness. The individual occasionally pays attention to details, exhibits poor organization skills, lacks discipline most times, inconsistently fulfills duties, and rarely meets commitments.*
>
> 3 *Demonstrates a lower than average level of conscientiousness. The individual sometimes pays attention to details, exhibits some organization, inconsistently shows discipline, performs duties sporadically, and occasionally fails to deliver on commitments.*
>
> 4 *Represents a slightly below average level of conscientiousness. The individual pays attention to details but may miss some, maintains a basic level of organization, shows discipline on an inconsistent basis, generally fulfills their duties, and often delivers on their commitments.*
>
> 5 *Represents an average level of conscientiousness. The individual pays reasonable attention to details, maintains organization that supports their tasks, demonstrates discipline occasionally, meets their duties most of the time, and usually delivers on their commitments.*
>
> 6 *Demonstrates a slightly above average level of conscientiousness. The individual usually pays attention to details, maintains good organization, often shows discipline, typically fulfills their duties, and usually meets commitments on time.*
>
> 7 *Shows a high level of conscientiousness. The individual consistently pays attention to details, maintains a well-structured organization, regularly exhibits discipline, fulfills their duties, and almost always meets commitments on time.*
>
> 8 *Represents a very high level of conscientiousness. The individual almost always pays meticulous attention to details, maintains excellent organization, exhibits strong discipline, is dutiful, and meets commitments on time.*
>
> 9 *Demonstrates an exceptional level of conscientiousness. The individual pays meticulous attention to details, maintains superior organization, shows exceptional discipline, strictly adheres to their duties, and consistently meets commitments ahead of schedule.*
>
> 10 *Represents the highest possible, extraordinary level of conscientiousness. The individual consistently demonstrates extreme attention to detail, maintains an impeccable level of organization, exhibits unwavering discipline, strictly adheres to their duties, and always delivers on their commitments ahead of schedule, often going beyond what is required.*

Because any one estimate by an expert human, or by a well-constructed prompt and LLM, may be highly uncertain (or worthless), a large array of these prompts are needed to produce the data for an analysis testing the viability of an objective measurement.

3.3.3 Quality assurance

While the potential for these new, alternative hypothesized formulations of measurands and constructs may be efficient and useful to collect raw data, it's equally essential that these new approaches conform to professional quality standards. Probabilistic models for measurement methodologically aligned with both the psychometric and metrological traditions (Mari & Wilson, 2014; Mari et al., 2023; Pendrill, 2014, 2019; Pendrill & Fisher, 2015) provide a robust framework for ensuring such quality.

Scientific measurement modeling demands that the prompts and AI algorithms used in measurement can be shown in theory and experiment to calibrate to invariant scale positions that function the same way across different groups of people or objects. This fundamental requirement informs the empirical testing of our newly formulated measurands or constructs, and contributes to demonstrating the instruments and inferences from them are robust, reliable, and valid across contexts and applications.

3.4 AI-driven remediation of distortions in measures

Even though these innovations portend a better metrological future, traditional pitfalls, such as insufficient sample sizes, can still pose significant challenges, even with the advent of AI measurement. Similarly, AI is notorious for its ability to be biased and could have both random and systematic errors that can distort measurements, leading to misinterpretations and incorrect conclusions. Bias is only noticeable in relation to a standard, and though standards in psychology and the social sciences are commonly assumed to be imposed on people and circumstances from the outside and from the top down, longstanding measurement theory and practice has amply demonstrated decades of results in which invariant patterns emerge from question-and-answer dialogues of their own accord. Over the last several decades, multiple instances of these patterns have persisted across many different samples of persons and assessment or survey items. The general structures obtained are already serving as standards in practical applications in several fields (Quaglia et al., 2016; Stenner, 2023).

At the same time, AI measurement offers avenues for improvement. The metrologically oriented paradigm for measurement allows us to automatically detect and adjust these distortions across the sciences. It helps to identify issues such as misfitting prompts or inconsistencies in synthetic raters by objectively estimating different parameters without respect to sample quirks that other social science approaches suffer from.

Additionally, multifaceted modeling approaches calibrate the severity or leniency biases of different measurement design facets affecting the measurement information, removing their effects from the estimates. These adjustments can be complemented by close attention to the details of response data, such that systematic biases for or against identifiable groups are brought to light and routinely reported to those affected and those responsible for their education or care (Fisher et al., 2021). By modeling the interactions between raters, items, and individuals, measurement stands as a clearly preferable alternative to summed scores that aids in identifying and managing biases or distortions. Measurement creates engineering-worthy levels of accuracy and precision that are taken for granted in metrology, but rare in the social sciences.

3.4.1 Hypothesized uses with physical metrology

While we don't have empirical evidence to report like the earlier examples, we suspect that LLM can also be useful for more traditional metrological tasks – either in direct measurement or improving intermediate steps. For example, the second author was involved in a project in 2023 at the Lawrence-Berkeley National Laboratory to try to create a measure of the electrical resistance of a material at different temperatures that are critical to superconductivity. Creating such a measure would help researchers understand the resistance loss and overall impurity in a superconductor that underlie other important engineering projects such as fusion-based power generation. One hypothesis was that one way to estimate the residual resistivity ratio (RRR) conveniently and inexpensively was to see if physical deformations of different parts of an intermetallic type II superconductor comprised of niobium and tin (Nb_3Sn) Rutherford cables could be used to estimate the resistance ratio in a cable. Having a good measure of this resistance loss would help researchers estimate losses and impurities in a cable much faster and far cheaper than trying to measure the loss in the entire cable or even a few samples of a cable.

A study in 2023 used images from existing Rutherford cables after passing current through each Nb_3Sn cable under very cold 4 Kelvin temperatures and allowing them to warm up to room temperature to calculate each samples' RRR value. To try to estimate a cheaper, faster, and more convenient instrument, the study took raw images using a reflective microscope, and then used a python script and Machine Learning algorithm (not an LLM) to determine the brightest parts of the image, creating a "mask" defining the brightest region of the image so that a machine learning algorithm could estimate the height, width, and area of the facet. The scientists directing the study were looking only for a good correlation between the more expensive RRR and the estimate of the new procedure, to see if it was a decent proxy.

Alternatively, this new LLM-based approach could significantly enhance the analysis. While the same procedures to prepare and analyze the samples would be required, a diverse array of prompts looking not only at the mask geometries used in the earlier study but also other hypothesized variables that could also provide information about the resistivity loss. By using a construct mapping approach to hypothesizing different elements of the samples' resistivity levels (zero loss, all the way to complete loss), there is the potential to get vastly better precision than only considering geometrical estimates. Further, because the LLM estimates can be subjected to a measurement modeling analysis, all the normal metrological quality assurances for linearity, accuracy, and precision can be estimated in addition to the usual correlations with results obtained using more expensive instruments. An example prompt that just replicates the 2023 study might be:

> *Please evaluate the provided image of a major axis facet of a Nb3Sn Rutherford cable, which has undergone either an acute or gradual bend, and will be operating under 4K Kelvin temperature with current flowing through it. Assume this is a standard Rutherford cable and the goal is to determine what, if any, impurities are present. Assess the impact on the Residual Resistivity Ratio (RRR) due to the deformation, using any visible microstructure alterations, and the physical dimensions (height and width) to calculate the cross-sectional area. Report this loss as a numerical value on a scale from 1 to 100.*

This approach would be familiar to a metrologist, as it mirrors the process they might go through in analyzing the data manually. However, by delegating this task to robotic workflow together with an AI system, we can potentially process larger volumes of measures more quickly, affordably, and consistently, while still maintaining a high degree of accuracy and reliability. What's unique about the use of a LLM for a metrologist is that different LLMs can examine different stimuli or signals, from slightly different approaches that can be evaluated later in relation to a probabilistic measurement model indicating whether the signals could be combined to reduce uncertainty, increase precision, reduce cost, or achieve other related goals.

3.4.2 Traceability to theory

A central aspect of robust measurement practice is traceability to theory. Measurements are not merely numbers; they are the operational definitions of the real-world interpretation of the theory and should be firmly anchored in the relevant science.

Automated measurement modeling plays a useful role in ensuring this type of traceability. By aligning item difficulty with the trait level of individuals, it operationalizes theoretical constructs. It provides a mechanism to verify whether the empirical observations correspond with theoretical expectations, a vital step in establishing the validity of any inferences made from these measures.

By automating a measurement analyst's judgments, this process liberates the measurement specialist to quickly and painlessly remediate distortions while ensuring measurements are meaningful, accurate, and grounded in theory. The incorporation of AI has revolutionized our ability to conduct such sophisticated analyses on a larger scale and at a faster pace, heralding a new era of more robust, reliable, and valid measurement practices.

3.5 Ethics

As we progress toward AI-driven transdisciplinary measurement, we must grapple with the ethical challenges presented by these advanced technologies. AI, while powerful, is not infallible. It can sometimes exhibit biases and "hallucinate," producing results that are not grounded in reality, and are even embarrassing. The implications of these issues are nontrivial, sparking a call by 33,000 academics, and other thought leaders like Elon Musk and Apple cofounder Steve Wozniak for a temporary freeze on AI development until we can implement more ethical and robust safeguards (Future of Life Institute, 2023).

A robust measurement approach that integrates metrological standards with industrial-organizational psychological techniques for AI safety may serve as an effective framework for addressing these issues (Barney & Fisher, 2017; Barney, 2019). However, the adoption of metrologically oriented approaches for AI "guardrails" remains limited within the industry. By making sure that properly calibrated instruments provide boundary conditions and ranges of appropriate action (e.g., by an AI or a human), we can improve the odds that the use of measurement information is appropriate. Implementing such approaches not only helps to prevent biases and hallucinations but also ensures the relevance and appropriateness of any interventions suggested by the AI. For instance, an AI-driven coaching system for swimming should be able to differentiate between a novice swimmer and an Olympic athlete. It would be potentially hazardous to suggest that a beginner jump into the deep end, just as it would be unhelpful to advise an Olympian to practice blowing bubbles in the shallow end. In essence, when measurements are going to be used by people to develop, feedback and coaching must be appropriately matched to the person's measured proficiency level, a task that can only be accomplished with careful, metrologically guided AI that is informed by the science about those people's tasks.

Ethics and persuasion are important to the measurement sciences, to make sure our instruments are used for good and to make sure that AI also promotes human wellbeing. Metrology – quality-assured measurement – has developed hand-in-hand with the wider quality assurance and conformity assessment and regularity communities. Metrological requirements are central to quality-assurance norms (written standards) such as the well-known ISO 9,000 series. This symbiosis between metrology and standards is expected to continue: For instance, new standards for data quality assurance are currently under development by the ISO/IEC JTC 1/SC 42 *Artificial intelligence* technical committee. Drawing from Cialdini's principles of persuasion, we find that effective and ethical measurement information must be produced in such a way that people using measurements understand what is truthful (bounded by uncertainty), naturally present (not contrived), and wise (promotes long-term relationships). As we delve deeper into the integration of AI and metrology, these principles become even more critical, particularly in the context of high-context cultures, where physical proximity and nonverbal behaviors are vital indicators of honesty and wisdom.

3.6 The future: automated prompts for transdisciplinary integration

The advent of automated prompts and multimodal AI brings an exciting new perspective to the future of transdisciplinary measurement. Not only do these technologies allow for the seamless incorporation of diverse raw data inputs for accurate, precise, linear, and practical measurements, but they also provide a unique liberating aspect. They free the measurement specialists, both metrologists and psychometricians, from the constraints traditionally associated with the measurement process.

With a prompt-based approach, the creative thinking of measurement professionals is no longer hampered by logistical, cost, organizational political considerations or by the face validity of the stimulus or costs of raw materials used to create potential measures. Measurement specialists are now free to explore, hypothesize, and innovate. We can craft prompts with varying levels of detail and nuance, explore competing hypotheses, and generate transdisciplinary measurement models that would be impossible with traditional approaches.

Moreover, improved automation of real-time measurement that was previously impossible opens up many new possibilities for what used to only be possible in high-tech semiconductor factories. The potential to automate measurement productivity in both industrial and social sciences marks the potential for a massive increase in the utility of all forms of measurement, when combined to realize important goals such as with Pritchard's ProMES approach to productivity improvement (Pritchard, Weaver & Ashwood, 2012) and the sophisticated tools of industrial engineering (e.g., statistical process control), and operations research to monitor and improve a wide variety of stochastic engineering, scientific and organizational systems.

The future of transdisciplinary measurement also lies in the integration of IoT technology. With the ability to gather and transmit vast amounts of diverse raw data, IoT devices can offer additional layers of depth and precision to measurement processes for human and engineering systems.

In essence, the fusion of these technologies with the rigor of existing and potential new metrological standards signifies a remarkable leap forward in our ability to measure and understand the world around us. By liberating the creative potential of measurement specialists and leveraging the power of advanced technology, we are entering a new era of measurement that is not only efficient and precise but also ethically sound and accountable.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23.

Allen, D. D., & Pak, S. (2023). Improving clinical practice with person-centered outcome measurement. In W. P. Fisher Jr. & S. J. Cano (Eds.), *Person centered outcome metrology* (pp. 53–105). Springer.

Andersen, E. B. (1999). Sufficient statistics in educational measurement. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 122–125). Pergamon.

Andrich, D. (1988). *Sage University Paper Series on Quantitative Applications in the Social Sciences. Vol. series no. 07–068: Rasch models for measurement.* Sage Publications.

Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika*, *75*(2), 292–308.

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, *37*(9), 387–416.

Andrich, D., & Hagquist, C. (2015). Real and artificial differential item functioning in polytomous items. *Educational and Psychological Measurement*, *75*(2), 185–207.

Anthony, C. J., Styck, K. M., Volpe, R. J., & Robert, C. R. (2023). Using many-facet Rasch measurement and generalizability theory to explore rater effects for direct behavior rating–multi-item scales. *School Psychology*, *38*(2), 119.

Attali, Y. (2018). Automatic item generation unleashed: An evaluation of a large-scale deployment of item models. In *International Conference on Artificial Intelligence in Education* (pp. 17–29). Springer.

Barney, M. (2019). The reciprocal roles of artificial intelligence and industrial-organizational psychology. In R. N. Landers (Ed.), *The Cambridge handbook of technology and employee behavior* (pp. 38–56). Cambridge University Press. https://doi.org/10.1017/9781108649636

Barney, M., & Fisher, W. P., Jr. (2016). Adaptive measurement and assessment. *Annual Review of Organizational Psychology and Organizational Behavior*, *3*, 469–490. https://www.annualreviews.org/doi/abs/10.1146/annurev-orgpsych-041015-062329

Barney, M., & Fisher, W. (2017). Avoiding AI Armageddon with metrologically-oriented psychometrics. *18th International Congress of Metrology*, *09005*, 1–6. https://doi.org/10.1051/metrology/201709005

Bateson, G. (1978). Number is different from quantity. *CoEvolution Quarterly*, 17, 44–46 [Reprinted from pp. 53–58 in Bateson, G. (1979). *Mind and nature: A necessary unity*. New York: E. P. Dutton.]. http://www.wholeearth.com/issue/2017/article/295/number.is.different.from.quantity

Bejar, I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning, and Assessment*, *2*(3), 1–29; http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1663.

Berglund, B., Rossi, G. B., Townsend, J., & Pendrill, L. R., (Eds.) *Measurements with persons: Theory, methods, and Implementation Areas*, Psychology Press, Taylor & Francis, 2013, ISBN 978-1-1367237-3-5

Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. arXiv preprint arXiv:2304.04616.

Blok, A., Farias, I., & Roberts, C. (Eds.). (2020). *The Routledge companion to actor-network theory*. Routledge.

Bohlig, M., Fisher, W. P., Masters, G. N., & Bond, T. (1998). Content validity and misfitting items. *Rasch Measurement Transactions*, *12*(1), 607 [http://www.rasch.org/rmt/rmt121f.htm].

Bowker, G., Star, S. L., Gasser, L., & Turner, W. (Eds.). (2014). *Social science, technical systems, and cooperative work: Beyond the great divide*. Psychology Press.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. University of Texas at Austin. https://www.liwc.app

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of pair comparisons. *Biometrika*, *63*, 324–345.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712v5. https://doi.org/10.48550/arXiv.2303.12712

Bullot, N. J., Seeley, W. P., & Davies, S. (2017). Art and science: A philosophical sketch of their historical complexity and codependence. *The Journal of Aesthetics and Art Criticism*, *75*(4), 453–463. doi:10.1111/jaac.12398

Cialdini, R. (2009). *Influence: Science and practice (5th ed.)*. Allyn & Bacon. ISBN-13 978-0-205-60999-4

Cialdini, R. (2018). *Pre-Suasion: A revolutionary way to influence and persuade*. Simon & Schuster. ISBN-13 978-1-501-10981-2.

Cialdini, R. (2021). *Influence: The psychology of persuasion* (revised and expanded ed.). HarperCollins. ISBN-13 978-0-061-89987-4.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. Psychological Science, 3, 186–190.

Comite International des Poids et Mesures (CIPM). (1999). Mutual recognition of national measurement standards and of calibration and measurement certificates issued by National metrology institutes. CIPM (Sevres: BIPM) (Technical Supplement revised in October 2003). www.bipm.org/en/cipm-mra/

Daniel, R. C., & Embretson, S. E. (2010). Designing cognitive complexity in mathematical problem-solving items. *Applied Psychological Measurement*, *34*(5), 348–364.

David, M. C. B., Kolanko, M., Del Giovane, M., Lai, H., True, J., Beal, E., Li, L. M., Nilforooshan, R., Barnaghi, P., Malhotra, P. A., Rostill, H., Wingfield, D., Wilson, D., Daniels, S., Sharp, D. J., & Scott, G. (2023). Remote monitoring of physiology in people living with dementia: An observational cohort study. *JMIR Aging*, *6*. https://doi.org/10.2196/43777

Dawson, T. L., Fischer, K. W., & Stein, Z. (2006). Reconsidering qualitative and quantitative research approaches: A cognitive developmental perspective. *New Ideas in Psychology*, *24*, 229–239.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Statistics for Social and Behavioral Sciences. Springer-Verlag.

Douglas, M. (1986). *How institutions think*. Syracuse University Press.

Dove, G., & Fayard, A. L. (2020). Monsters, metaphors, and machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). ACM. https://doi.org/10.1145/3313831.3376275

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380–396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*(4), 407–433.

Embretson, S. E. (2010). *Measuring psychological constructs: Advances in model-based approaches*. American Psychological Association.

Evans, K. (1996). Review: Chaos as opportunity: Grounding a positive vision of management and society in the new physics. *Public Administration Review*, *56*(5), 491–494.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, *46*(1), 59–77.

Fisher, W. P., Jr. (2009). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement*, *42*(9), 1278–1287.

Fisher, W. P., Jr. (2010). The standard model in the history of the natural sciences, econometrics, and the social sciences. *Journal of Physics Conference Series*, *238*(1), http://iopscience.iop.org/1742-6596/238/1/012016/pdf/1742-6596_238_1_012016.pdf.

Fisher, W. P., Jr. (2022). Aiming higher in conceptualizing manageable measures in production research. In N. Durakbasa & M. G. Gençyilmaz (Eds.), *Digitizing production systems: Selected papers from ISPR2021*, *October 07–09*, *2021 online, Anatalya, Turkiye* (pp. xix–xxxix). Springer Verlag. https://link.springer.com/content/pdf/bfm%3A978-3-030-90421-0%2F1

Fisher, W. P., Jr. (2023). Separation theorems in econometrics and psychometrics: Rasch, Frisch, two Fishers, and implications for measurement. *Journal of Interdisciplinary Economics*, *35*(1), 29–60. https://journals.sagepub.com/doi/10.1177/02601079211033475

Fisher, W. P., Jr., & Cano, S. (Eds.). (2023). *Person-centered outcome metrology: Principles and applications for high stakes decision making*. Springer Series in Measurement Science & Technology. Springer. https://link.springer.com/book/10.1007/978-3-031-07465-3

Fisher, W. P., Jr., Oon, E. P.-T., & Benson, S. (2021). Rethinking the role of educational assessment in classroom communities: How can design thinking address the problems of coherence and complexity? *Educational Design Research*, *5*(1), 1–33

Fisher, W. P., Jr, & Stenner, A. J. (2013). On the potential for improved measurement in the human and social sciences. In Q. Zhang & H. Yang (Eds.), *Pacific Rim Objective Measurement Symposium 2012 Conference Proceedings* (pp. 1–11). Springer-Verlag. https://link.springer.com/chapter/10.1007/978-3-642-37592-7_1

Fisher, W. P., Jr., & Stenner, A. J. (2016). Theory-based metrological traceability in education: A reading measurement network. Measurement, 92, 489–496. [Reprinted in W. P. Fisher Jr. & P. J. Massengill (Eds.) (2023). *Explanatory models, unit standards, and personalized learning in educational measurement: Selected papers by A. Jackson Stenner*, (pp. 275–293). Springer. https://link.springer.com/book/10.1007/978-981-19-3747-7]

Future of Life Institute. (2023). Pause giant AI experiments: An open letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Gehrmann, S., Clark, E., & Sellam, T. (2022). Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. arXiv preprint arXiv:2202.06935.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, *59*, 167–192. https://doi.org/10.1146/annurev.psych.58.110405.085632

Guttman, L. (1971). Measurement as structural theory. *Psychometrika*, *36*, 329–347. https://doi.org/10.1007/BF02291362

Habermas, J. (1995). *Moral consciousness and communicative action*. MIT Press.

Haraway, D. J. (1992). The promises of monsters: A regenerative politics for inappropriate/d others. In L. Grosberg, C. Nelson, & P. Treichler (Eds.), *Cultural Studies* (pp. 295–336). Routledge.

Haraway, D. J. (2022). A Cyborg Manifesto: An ironic dream of a common language for women in the integrated circuit. In *The Transgender studies reader remix* (pp. 429–443). Routledge.

Hayman, J., Rayder, N., Stenner, A. J., & Madey, D. L. (1979, August 1). On aggregation, generalization, and utility in educational evaluation. *Educational Evaluation and Policy Analysis*, *1*(4), 31–39.

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement*, *10*(4), 369–380. *Review of Psychology*, *59*, 167–192. https://doi.org/10.1146/annurev.psych.58.110405.085632

Gierl, M. J., & Haladyna, T. M. (2012). *Automatic item generation: Theory and practice*. Routledge.

HuggingFace. (2023). Open LLM leaderboard. Hugging face. Retrieved May 29, 2023, from https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Jasanoff, S. (2004). *States of knowledge: The co-production of science and social order*. International Library of Sociology. Routledge.

JCGM. (2020). Guide to the expression of uncertainty in measurement – Part 6: Developing and using measurement models. https://www.bipm.org/documents/20126/50065290/JCGM_GUM_6_2020.pdf

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kosh, A., Simpson, M. A., Bickel, L., Kellog, M., & Sanford-Moore, E. (2019). A cost-benefit analysis of automatic item generation. *Educational Measurement: Issues and Practice*, *38*(1), 48–53.

Kristeva, J. (2012). *The severed head: Capital visions*. Columbia University Press.

Latour, B. (2012). Love your monsters: Why we must care for our technologies as we do our children. *Breakthrough Journal*, *2*, 21–28. Retrieved 1 January 2016, from https://thebreakthrough.org/journal/issue-2/love-your-monsters.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Sage Publications. ISBN 0-941938-02-6

Linacre, J. M. (2003). Size vs. significance: Infit and outfit mean-square and standardized chi-square fit statistics. *Rasch Measurement Transactions*, *17*(1), 918 [http://www.rasch.org/rmt/rmt171n.htm].

Linacre, J. M. (2022). *A user's guide to FACETS Rasch-Model computer program, v. 3.84.1*. Winsteps.com. http://www.winsteps.com/a/facets-manual.pdf

Linacre, J. M., & Wright, B. D. (1994). Dichotomous infit and outfit mean-square fit statistics. *Rasch Measurement Transactions*, *8*(2), 350 [http://www.rasch.org/rmt/rmt82a.htm].

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, *61*(4 (Whole No. 285)), 1–49.

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review*, *72*(2), 143–155.

Luce, R. D. (1959). *Individual choice behavior*. Wiley.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new kind of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27.

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement*, *51*, 315–327. http://www.sciencedirect.com/science/article/pii/S0263224114000645

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement*. 2nd Ed. Springer Nature. https://doi.org/10.1007/978-3-030-65558-7

Massof, R. W., & Bradley, C. (2023). An adaptive strategy for measuring patient-reported outcomes. In W. P. Fisher Jr. & S. J. Cano (Eds.), *Person-centered outcome metrology* (pp. 107–150). Springer.

Merry, S. E. (2016). *The seductions of quantification: Measuring human rights, gender violence, and sex trafficking*. University of Chicago Press.

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, *99*(2), 166–180.

Norris, C. (1999). Sexed equations and vexed physicists: The 'Two Cultures' revisited. *International Journal of Cultural Studies*, *2*, 77–107.

Patil, S. G., Zhang, T., Wang, X., & Gonzalez, J. E. (2023). Gorilla: Large Language Model connected with massive APIs. https://arxiv.org/abs/2305.15334

Pendrill, L. (2019). *Quality assured measurement: Unification across social and physical sciences*. Springer Nature. https://doi.org/10.1007/978-3-030-28695-8

Pendrill, L. (2023) Quantities and units: Order amongst complexity. In W. P. Fisher Jr. & L. Pendrill, *Models, measurement, and metrology extending the SI*. De Gruyter.

Pendrill, L. R. (2014). Man as a measurement instrument [Special Feature]. NCSLi Measure: The Journal of Measurement Science, 9(4), 22–33. http://www.tandfonline.com/doi/abs/10.1080/19315775.2014.11721702

Pendrill, L., & Fisher, W. P., Jr. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement*, *71*, 46–55. http://dx.doi.org/10.1016/j.measurement.2015.04.010

Poinstingl, H. (2009). The linear logistic test model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychology Science Quarterly*, *51*, 123–134.

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Postman, N. (1992). *Technopoly: The surrender of culture to technology*. Vintage Books.

Power, M. (2004). Counting, control and calculation: Reflections on measuring and management. *Human Relations*, *57*, 765–783.

Pritchard, R. D., Weaver, S. J., & Ashwood, E. L. (2012). *Evidence-based productivity improvement: A practical guide to the productivity measurement and enhancement system (ProMES)*. Routledge/Taylor & Francis Group. https://doi.org/10.4324/9780203180341

Quaglia, M., Pendrill, L., Melin, J., & Cano, S., & 15HLT04 NeuroMET Consortium. (2016–2019). *Innovative measurements for improved diagnosis and management of neurodegenerative diseases (EMPIR NeuroMET)*. Teddington, Middlesex, UK: EURAMET. https://www.lgcgroup.com/our-programmes/empir-neuromet/neuromet-landing-page/ (36 pp.)

Quaglia, M., Pendrill, L., Melin, J., & Cano, S., & 18HLT09 NeuroMET2 Consortium. (2019–2022). *Metrology and innovation for early diagnosis and accurate stratification of patients with neurodegenerative diseases (EMPIR NeuroMET)*. Teddington, Middlesex, UK: EURAMET. https://www.lgcgroup.com/our-programmes/empir-neuromet/neuromet-landing-page/ (5 pp.)

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Reprint, with Foreword and Afterword by B. D. Wright. University of Chicago Press, 1980). Danmarks Paedogogiske Institut.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability: Vol. IV: Contributions to biology and problems of medicine* (pp. 321–333 [http://www.rasch.org/memo1960.pdf]). University of California Press.

Rousseau, D. M. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. *Research in Organizational Behavior*, *7*(1), 1–37.

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., Christiano, P., & Dafoe, A. (2023). Model evaluation for extreme risks. arXiv preprint arXiv:2305.15324.

Snow, C. P. (1964). *The two cultures and a second look*. Cambridge University Press.

Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quarterly*, *50*(3), 345–362.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. Information Systems Research, 7(1), 111–134.

Staudinger, U. M., & Glück, J. (2011). Psychological wisdom research: Commonalities and differences in a growing field. *Annual Review of Psychology*, *62*, 215–241. https://doi.org/10.1146/annurev.psych.121208.131659

Stenner, A. J. (2023). Measuring reading comprehension with the Lexile Framework. In W. P. Fisher Jr. & P. J. P. J. Massengill (Eds.), *Explanatory models, unit standards, and personalized learning in educational measurement* (pp. 63–88). Springer.

Stenner, A. J., Fisher, W. P., Jr, Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology: Quantitative Psychology and Measurement*, *4*(536), 1–14. doi: 10.3389/fpsyg.2013.00536

Stone, M. H., Wright, B., & Stenner, A. J. (1999). Mapping variables. *Journal of Outcome Measurement*, 3(4), 308–322. http://jampress.org/JOM_V3N4.pdf. (Rpt., W. P. Fisher Jr. & P. J. Massengill (Eds.), (2023). *Explanatory models, unit standards, and personalized learning in educational measurement: Selected papers by A. Jackson Stenner* (pp. 109–120). Springer. https://link.springer.com/chapter/10.1007/978-981-19-3747-7_8)

Tal, E. (2014). Making time: A study in the epistemology of measurement. *The British Journal for the Philosophy of Science*, *67*(1), 297–335. https://doi.org/10.1093/bjps/axu037

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, XXXIII, 529–544. (Rpt. in L. L. Thurstone, (1959), *The measurement of values* (pp. 215–233). University of Chicago Press, Midway Reprint Series.)

Wilmot, M. P., & Ones, D. S. (2019). A century of research on conscientiousness at work. *Proceedings of the National Academy of Sciences*, *116*(46), 23004–23010. https://doi.org/10.1073/pnas.1908430116

Wilson, M. R. (Ed.). (2004). *National Society for the Study of Education Yearbooks. Vol. 103, Part II: Towards coherence between classroom assessment and accountability.* University of Chicago Press.

Wilson, M. (2023). *Constructing measures: An item response modeling approach, 2nd ed*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410611697

Wilson, M., & Gochyyev, P. (2020). Having your cake and eating it too: Multiple dimensions and a composite. *Measurement, 151*(107247).

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97–116 http://www.rasch.org/memo42.htm

Wright, B. D. (1994). Theory construction from empirical observations. *Rasch Measurement Transactions*, *8*(2), 362. http://www.rasch.org/rmt/rmt82h.htm

Wright, B. D. (1996). Reliability and separation. *Rasch Measurement Transactions*, *9*(4), 472 [http://www.rasch.org/rmt/rmt94n.htm].

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*(4), 33–45, 52. https://doi.org/10.1111/j.1745-3992.1997.tb00606.x

Wright, B. D. (2000/2008). Steps leading to a straight line: Constructing a variable [Psychometric Theory class handout, University of Chicago, 2000]. *Rasch Measurement Transactions*, 22(1), 1155.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.

Wright, B. D., & Stone, M. H. (2003). Five steps to science: Observing, scoring, measuring, analyzing, and applying. *Rasch Measurement Transactions*, *17*(1), 912–913 [http://www.rasch.org/rmt/rmt171j.htm].

Xie, Y., & Wilson, M. (2008). Investigating DIF and extensions using an LLTM approach and also an individual differences approach: An international testing context. *Psychology Science Quarterly*, *50*, 403–416.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., et al. (2023). A survey of large language models. ArXiv Preprint ArXiv:2303.18223.

Ernesto San Martín, Marcela Perticará, Inés M. Varas,
Kenzo Asahi, and Jorge González

4 The role of identifiability in empirical research

Abstract: This chapter discusses the general concepts of identification and partial identification of statistical models. We elucidate the identification restrictions to endow with meaning the parameters of interest of the fixed-effects one-parameter logistic model with guessing (1PL-G), a model used in educational measurement. We also review the restrictions for identifying the average treatment effect (ATE) in evaluating a policy or program. To address the fundamental problem of causal inference, we also present a partial identification analysis of the ATE. On the basis of the results, we emphasize the relevance of an identification analysis and the usefulness of considering a partial identification approach in causal inference.

Keywords: Partial identification, Causal inference, Self-selection bias, Finite sample space

> [. . .] it has been considered legitimate to use some
> of the *tools* developed in statistical theory *without*
> accepting the very *foundation* upon which
> statistical theory is built.
> (Haavelmo, 1944)

4.1 Introduction

Manski (2013a) emphasizes combining assumptions and data to draw meaningful conclusions in empirical research and policy analysis. The logic of empirical inference can accordingly be summarized as follows:

Acknowledgments: The authors were partially supported by the Millennium Nucleus on Intergenerational Mobility MOVI NCS2021072. Kenzo Asahi and Jorge González were also partially funded by Chile's Agencia Nacional de Investigación y Desarollo (ANID)'s FONDECYT Regular 1231755 and 1230968 respectively.

Ernesto San Martín, Inés M. Varas, Jorge González, Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI), Chile; Interdisciplinary Laboratory of Social Statistics LIES, Pontificia Universidad Católica de Chile; Faculty of Mathematics, Pontificia Universidad Católica de Chile
Marcela Perticará, Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI), Chile; Facultad de Administración y Economía, Universidad Diego Portales, Chile
Kenzo Asahi, Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI), Chile; Escuela de Gobierno, Pontificia Universidad Católica de Chile; Centre for Sustainable Urban Development (CEDEUS), Chile ORCID ID https://orcid.org/0000-0001-7838-4647

https://doi.org/10.1515/9783111036496-004

$$\text{Assumptions} + \text{Data} \Rightarrow \text{Conclusions}$$

Merely relying on data alone is insufficient for making conclusions; researchers need to make explicit assumptions linking the data to the population of interest.

It is essential to distinguish between these assumptions and the statistical hypotheses that can be falsified with data through the “null ritual” of hypothesis testing (Gigerenzer, 2004). The former are the so-called maintained assumptions, which makes explicit a fundamental difficulty of empirical research: holding fixed the available data, what assumptions can we maintain (Manski et al., 2021)? These considerations allow us to distinguish between data and (scientific) knowledge: “evidence is synonymous with data. Knowledge is the set of conclusions that one draws by combining evidence with assumptions about unobserved quantities” (Manski, 2013b).

Throughout this chapter, we pursue Manski’s perspective by relating those assumptions on unobserved quantities with the concept of *parameter of interest*. To make this link explicit, we first establish what we understand by *population* and *structure*, and second, we clarify the importance of distinguishing and defining the identified parameters and the parameters of interest (Section 4.2).

After that, in Section 4.3, we present the identification analysis of models commonly used in educational measurement and psychometrics to represent the underlying response process of test takers to a set of stimuli/items (Lord, 1980; Hambleton & Swaminathan, 1985).

Section 4.4 provides a detailed discussion of the identification problem in causal inference, mainly focusing on the average treatment effect (ATE) in an observational study. We argue that causal inference often relies on strong assumptions and additional information; it is crucial to carefully interpret the estimated causal effects based on the underlying model and data limitations. We also explore the possibility of identifying the parameters of interest using logically weaker identification constraints.

In Section 4.5, we continue the discussion on the partial identification of the parameters of interest using a specific example of self-selection in the context of a leveling program of a Chilean public university. Some conclusions are gathered in Section 4.6.

4.2 The logic of empirical research

4.2.1 Population and structure

Following Koopmans & Reiersol (1950) and Hurwicz (1950), the identification problem arises from distinguishing between a *population* “in the sense of a [probability] distribution of observable variables” and a *structure* referring to the “investigator’s ideas regarding the explanation or formation of the phenomena studied.” In this way, Koopmans and Reiersol (1950) reformulate the *specification problem* originally made explicit by Fisher (1922). Instead of specifying the "mathematical form of the probability distribution of the population," one specifies a set of structures corresponding to the probability distribution of certain unobserved variables and a specific relationship between observed and unobserved variables.

This framework gives rise to a fundamental problem, specifically, the question of whether a single structure uniquely generates the probability distribution of the observations. It is possible to focus on observations, seeking empirical relationships likely to be caused by the presence and persistence of underlying structural relationships. "However, the direction of this deduction cannot be reversed –from empirical to the structural relationships– except possibly with the help of a theory which specifies the form of the structural relationships, the variables which enter into each, and any further details supported by prior observation or deduction therefrom" (Koopmans, 1949). When multiple structures are compatible with the empirical evidence, it becomes challenging to assess their validity objectively. This situation leads to undesirable implications. At the policy level, different courses of action may be justifiable based on the same empirical evidence. At the scientific level, different conclusions stemming from the same empirical evidence might be considered as equally valid "scientific knowledge." This situation is frustrating: "all you can do is judge the persuasiveness of the arguments offered. If you are persuaded by one social scientist more than by another, it is only because one is a more skilled advocate for his or her position" (Manski, 1995, p. 2).

The previous framework involves the specification of (i) the latent variable model, say $p^{\omega_1}(\theta)$, where ω_1 is the corresponding parameter, and (ii) the conditional model (i.e., the structural relationships between observed and latent variables) of the observed data given the latent ones, say $p^{\omega_2}(y \mid \theta)$, where ω_2 is the corresponding parameter. Both models induce a probability distribution of the observable variables, namely

$$p^{(\omega_1,\omega_2)}(y) = \int p^{\omega_2}(y \mid \theta)p^{\omega_1}(\theta)d\theta \tag{4.1}$$

The parameters ω_1 and ω_2 constitute the essential component of the structure underlying the observed variables. Once these parameters are identified, we obtain a unique structural explanation.

4.2.2 Identified parameters and parameters of interest

Although Koopmans and Reiersol's framework (1950) may seem, at first glance, applicable only to structural models with latent variables, it is possible to highlight from (4.1) the fundamental objective underlying such a framework, namely "to learn what conclusions can and cannot be drawn given specified combinations of assumptions and data" (Manski, 1995, p. 3).

In the first level of analysis, the attention can be focused on a description of a set of data – the population of interest. What we *can learn from the data* corresponds to its probability distribution: its specification “is not arbitrary but requires an understanding of the way in which the data are supposed to, or did in fact, originate” (Fisher, 1973, p. 8). Parameters always index such probability distribution: these parameters exhaustively describe the population of interest “in respect of all qualities under discussion” (Fisher, 1973). This initial level of analysis, accordingly, requires making explicit three components:

1. The set of observations, along with the events of interest. This is the so-called sample space, which corresponds to a measurable space $(M, \mathcal{M})$ corresponding to the statistical units’ labels and constitutes a finite set (Basu, 1977). Consequently, the σ-field $\mathcal{M}$ reduces to the class of all the subsets of M.
2. A probability distribution P^a that is defined on the sample space and indexed by a parameter a. We refer to this probability distribution as *sampling probability* emphasizing its relation to the observed data.
3. The set A of all logical ranges of the parameter a, namely the *parameter space.*

These three components define the *statistical model*, which can be compactly written as

$$\varepsilon = \{(M, \mathcal{M}), P^a : a \in A\} \tag{4.2}$$

and that corresponds to a family of sampling probability distributions defined on the observed data, following Koopmans & Reiersol (1950) and Hurwicz (1950), Gourieroux and Monfort (1995; Chapter 1); McCullagh (2002); Florens et al. (2007, Chapter 1). In ε, the parameter a is said to be identified if the mapping Φ from the parameter space A into the set $P(M, \mathcal{M})$ of sampling probabilities defined on $(M, \mathcal{M})$ is such that $\Phi\,(a) = P^a$ is injective.

Remark 4.2.2 It should be noted that (4.2) involves the so-called parametric, nonparametric, and semi-parametric models. The parametric models are characterized by the fact that the parameter space A is (a subset of) a finite-dimensional vector space. For the nonparametric models, A corresponds to (a subset of) an infinite-dimensional vector space, for instance, a functional space or a space of probability distributions. Finally, for the semi-parametric model, the set A corresponds to a Cartesian product between a finite and infinite vector space.

It can be shown that a statistical model always involves an identified parameter; see Florens et al. (1985) and Florens et al. (1990, chapter 4). This suggests specifying the statistical model (4.2) in an identified way, namely, to index the sampling probabilities by the identified parameter, say $h(a)$, where h is a function defined from A into $h(A)$. Therefore, this identified parameterization fully captures the characteristics of the population of interest that can be gathered at this first level of analysis.

However, empirical research usually develops a second level of analysis. Faced with a set of observations, a researcher asks substantive questions. For example, in the second round of a presidential election, we only observe the proportion of votes in favor of one candidate or the other. If the difference in votes is minimal, a substantive question is whether there was electoral fraud. A political scientist would *like to learn from the data* whether there was electoral fraud. However, considering the observed data, such a political phenomenon is unobserved. The modeling challenge is, therefore, to express the electoral fraud as a parameter, which we will call a *parameter of interest*. The identification problem arises when the parameter of interest cannot be expressed as an injective function of the identified parameter. That is, when what we can learn from the data does not match what we want to know.

Technically speaking, a parameter of interest is a function of a, namely $g(a)$, where g is a function defined from A into $g(A)$; see Engle et al. (1983). Therefore, $g(a)$ is identified if the mapping Ψ from $g(A)$ into $P(M, \mathcal{M})$ such that $\Psi\,(\mathrm{g}(a)) = P^a$ is injective; see LeCam & Schwartz (1960) and Mouchart and Oulhaj (2006).

A strategy of identification analysis consists of establishing in (4.2) an injection Λ between the parameter of interest $g(a)$ and the identified parameter $h(a)$. Since only the identified parameters capture properties of the population under study, this strategy shows what needs to be established so that *what is to be learned from the data* matches *what can be learned from it*: that the mapping Λ is injective. Technically speaking, this strategy is based on the following commutative diagram:

$$\begin{tikzcd}
g(A) \arrow[r, "\Lambda\text{ injective}"] \arrow[dr, "\Psi = \Lambda \circ \Phi\text{ injective}"'] & h(A) \arrow[d, "\Phi\text{ injective}"] \\
& P(M, \mathcal{M})
\end{tikzcd}$$

This strategy will be pursued in the examples discussed in the following sections.

4.3 Item response theory models: What do difficulty and guessing parameters mean?

Item response theory (IRT) models are typically used to analyze a set of binary data related to the reaction/response of test takers to a set of stimuli/items. More precisely, the investigator is confronted with a matrix of dimension $I \times J$, whose entries are 0's and 1's: I corresponds to the number of persons, whereas J corresponds to the number of stimuli or items; an entry 0 corresponds either to a negative reaction of a person to a stimulus or to an incorrect response of a person to an item; an entry 1 corresponds to a positive reaction of a person to a stimulus or to a correct response of a person to an item. The $I \times J$ matrix can accordingly be represented as

$$\begin{pmatrix} y_{11} & \cdots & y_{1J} \\ \vdots & \vdots & \vdots \\ y_{I1} & \cdots & y_{IJ} \end{pmatrix}, \tag{4.3}$$

where $y_{mj} \in \{0,1\}$ for each $(m,j) \in \{1,\ldots,I\} \times \{1,\ldots,J\}$; here y_{mj} corresponds to the reaction or response of a person m to stimuli/item j. Therefore, the available information consists not only of the response patterns but also of a set of labels representing those people and a set of labels representing the stimuli/items. Thus, the available data are assembled into a set whose elements are ordered triplets (label of a person, label of a stimulus/item, response pattern):

$$\left\{(1,1,y_{11}),(1,2,y_{12}),\ldots,\left(1,J,y_{1J}\right),\ldots,(I,1,y_{I1}),(I,2,y_{I2}),\ldots,\left(I,J,y_{IJ}\right)\right\}$$

for a similar way of describing the available data, see Bahadur et al. (2002).

To formalize the components of the statistical model, note that once the set of stimuli/items is defined, the observed data comes only from the people exposed to such a set of stimuli/items. Thus, the sample space is defined as the set of labels associated with each person, that is, $M = \{1,\ldots,I\}$. Taking into account that a function is fully characterized by its image space, it is possible to define a random variable (or a function) Y as follows:

$$Y: M \longrightarrow \{0,1\}^J$$

such that for all $m \in M$,

$$Y(m) = \left(y_{m1},\ldots,y_{mJ}\right)$$

represents the person's response pattern. Note that the function Y can also be written as

$$Y(m) = (X(m,1),\ldots,X(m,J))$$

where $X: M \longrightarrow \{0,1\}$ and, consequently, $X(m,j) = \left(\pi_j \circ Y\right)(m,(1,\ldots,J))$, with π_j defined as $\pi_j(1,\ldots,j,\ldots,J) = j$ (Itô, 1984). It is assumed that the random variables $\{X(m,j): m \in M, j \in \{1,\ldots,J\}\}$ are mutually independent, and that each $X(m,j)$ is distributed according to a Bernoulli distribution with parameter p_{mj}. To simplify notation, we will denote by $X_j(m)$ the random variable $X(m,j)$ for all $(m,j) \in M \times \{1,\ldots,J\}$.

The reader may wonder what is the advantage in specifying the response variable Y in this way compared to the standard form in the IRT model literature (Lord & Novick, 1968; De Boeck & Wilson, 2004). There are at least three advantages:

1. The sample space is made explicit considering no elements related to the items, thus emphasizing the uniqueness of the available data (Fisher, 1955). This establishes the limits of statistical inference, which, in principle, must be reduced to the sample space.

2. The behavior of each person exposed to each stimulus/item is described by *one and only one* random variable. This contrasts with Lord and Novick's representation (1968, Chapter 2), in which one random variable is defined for each pair person-stimulus/item.
3. The definition of a unique random variable for a person's response pattern clarifies that people are distinguished whenever they have different response patterns, that is, when people belong to different elements on the equivalence class defined by the random variable Y on M.

4.3.1 A fixed-effects logistic model

To demonstrate the significance of the question that introduces this section, let us briefly examine the problem of parameter interpretability in a fixed-effects model.

Following Rasch (1960), the substantive problem is comparing persons using their responses to stimuli/elements. This problem led Rasch to introduce two parameters, one representing a specific characteristic of a person, say $\epsilon_m \in \mathbb{R}_+$, and the other representing a particular characteristic of the stimulus/item, say $\eta_j \in \mathbb{R}_+$, such that

$$p_{mj} = P\big(X_j(m) = 1\big) = F\left(\frac{\epsilon_m}{\eta_j}\right), \quad m \in M, j = 1, \ldots, J \tag{4.4}$$

where $F(x) = \mathrm{x}/[1+\mathrm{x}]$ for $x \in \mathbb{R}_+$. It should be remarked that the parameters of interest are $\{(\epsilon_m, \eta_j) \colon (m,j) \in M \times \{1, \ldots, J\}\}$, whereas the identified parameters are $\{p_{mj} : (m,j) \in M \times \{1, \ldots, J\}\}$.

The substantive problem can be solved once the meaning of ϵ_m and η_j is made explicit. Since Lord and Novick (1968, chapter 17) (see also De Boeck and Wilson, 2004; Van der linden and Hambleton, 1997; Baker and Kim, 2004), the standard psychometric literature has made explicit the meaning of the parameters of interest using the item characteristic curves, which correspond to the conditional probability of the observable variable $X_j(m)$ given a latent variable (in this case, what they call the person's ability). However, because the parameters correspond to the characteristics of the population under study, their meaning must be established concerning the statistical model. In passing, this confusion in the psychometric literature is (almost) the same as the one seen in the econometric literature, where models with fixed and random effects are confused and even compared; for some examples, see Longford (2012); Castellano et al. (2014); Clarke et al. (2015); Bell et al. (2019).

As discussed in the previous section, the strategy to identify the parameters of interest consists of establishing a one-to-one relationship between them and the identified parameters. Such a relationship follows noticing in (4.4) that, for every pair (m, j),

$$\frac{\epsilon_m}{\eta_j} = F^{-1}\left(p_{mj}\right)$$

If there are at least two items and one person, it follows that

$$\epsilon_m = \eta_1 \cdot F^{-1}(p_{m1}) \text{ for all } m \in M, \quad \eta_j = \eta_1 \cdot \frac{F^{-1}(p_{m1})}{F^{-1}\left(p_{mj}\right)} \text{ for all } j = 2, \ldots, J \tag{4.5}$$

Thus, ϵ_m's and η_j's depend on η_1 and, consequently, a necessary and sufficient condition for identifying the parameters of interest is to fix η_1. If $\eta_1 = 1$, then the characteristic ϵ_m of a person m corresponds to the betting odd of a correct answer to the standard item 1, and the characteristic η_j of item j could be interpreted as an odd ratio between item 1 and item j for each person m; for details, see Rasch (1966) and San Martín et al. (2009). It is important to emphasize that these interpretations of the parameters of interest are not based on psychological or educational considerations.

4.3.2 A model with a guessing parameter

Considering the same data (4.3), researchers have wondered whether a person can answer an item or a stimulus by "guessing." This question is more pressing in educational measurement, particularly when a standardized test has no consequences for individuals. In Chapter 17 of Lord and Novick (1968), Birnbaum introduced a latent trait model that allows random guessing so that "subjects of very low ability will sometimes give correct responses to multiple-choice items, just by chance." Birnbaum emphasizes the substantive side of the problem:

> A highly schematised psychological hypothesis has suggested one model for such items. This model assumes that if an examinee has ability θ, then the probability that he will know the correct answer is given by a normal ogive function $\Phi\left[a_g\left(\theta - b_g\right)\right]$ [here, Φ is the cumulative distribution function of a standard normal distribution, whereas a_g and b_g are item parameters] [. . .]; it further assumes that if he does not know it he will guess, and, with probability c_g, will guess correctly. It follows from these assumptions that [. . .] the probability of a correct response is the item characteristic curve
>
> $$Q_g(\theta) = c_g + \left(1 - c_g\right)\Phi\left[a_g\left(\theta - b_g\right)\right]$$
>
> The psychological hypothesis implicit here has been mentioned primarily to point up a mathematical feature of this form; the empirical validity of this form is not dependent on this psychological hypothesis. (p. 404)

For Birnbaum, "answering an item correctly by chance" is formulated using a probability that depends only on the item and not on the person's characteristics. However, when interpreting this probability, Birnbaum does not do so concerning the statistical model but only based on a conditional, unobservable model.

To understand the meaning of a guessing parameter, we will focus on a slightly simplified model called the 1PL-G model (Weitzman, 1996; San Martin et al., 2006). The 1PL-G fixed-effects model is specified as follows:

$$P\big(X_j(m)=1\big)=c_j+\big(1-c_j\big)F(\theta_m-\beta_j) \quad \text{for all } (m,j)\in M\times\{1,\ldots,J\} \tag{4.6}$$

where θ_m is a person parameter, β_j and c_j are known as the difficulty and guessing parameter related to the item, respectively; $F(x)=\exp(x)/[1+\exp(x)]$ with $x\in\mathbb{R}$ and $(c_j,\beta_j,\theta_m)\in[0,1]\times\mathbb{R}\times\mathbb{R}$. It is also assumed that $\big\{X_j(m){:}(m,j)\in M\times\{1,\ldots,J\}\big\}$ are mutually independent.

Note that (4.6) can be rewritten as

$$q_{mj}=P\big(X_j(m)=0\big)=\delta_jG(\theta-\beta_j) \quad \text{for all } (m,j)\in M\times\{1,\ldots,J\} \tag{4.7}$$

where $\delta_j=1-c_j\in[0,1]$ and G is a function such that $G(x)+F(x)=1$ for all $x\in\mathbb{R}$.

Assuming that there are at least two persons and two items, (4.7) implies the following equations:

$$\theta_m=G^{-1}\left(\frac{q_{m1}}{\delta_1}\right)+\beta_1 \tag{4.8}$$

$$\beta_j=G^{-1}\left(\frac{q_{m1}}{\delta_1}\right)-G^{-1}\left(\frac{q_{mj}}{\delta_j}\right)+\beta_1 \tag{4.9}$$

$$G^{-1}\left(\frac{q_{1j}}{\delta_\text{j}}\right)-G^{-1}\left(\frac{q_{2j}}{\delta_j}\right)=G^{-1}\left(\frac{q_{11}}{\delta_1}\right)-G^{-1}\left(\frac{q_{2j}}{\delta_j}\right) \tag{4.10}$$

Thus, $\theta_m=\theta_m(\beta_1,\delta_1)$, $\beta_j=(\beta_1,\delta_1)$, and $\delta_j=\delta_j(\delta_1)$. Therefore, a necessary and sufficient condition for identifying the parameters of interest is to fix the item parameters of an item, namely $(\beta_1,\delta_1)=(0,1)$ or, equivalently, $(\beta_1,c_1)=(0,0)$. This restriction reveals that there is no other way to know about the guessing parameter of the items than when there is at least one item with a guessing parameter equal to zero. Details about these results can be found in Appendix A.

These equations allow us to interpret the parameters of interest of the 1PL-G fixed-effect model:

1. Regarding the person parameter θ_m, it can be verified that its meaning is the same as that in an identified Rasch fixed-effects model. Moreover,

$$\theta_m>\theta_l \Leftrightarrow P(X_1(l)=1)<P(X_1(m)=1),$$

which provides empirical insight.

2. The item parameter β_j does not have the same meaning as in the identified Rasch fixed-effect model, which implies that comparing these parameters from one model to another is incorrect. Moreover,

$$\beta_j > \beta_k \Leftrightarrow \frac{q_{mj}}{\delta_j} > \frac{q_{mk}}{\delta_k}.$$

Thus, the sentence *item j* is *more difficult than item k* needs to be understood in the following terms: the probability of answering item j incorrectly is greater than the probability of answering item k incorrectly once both probabilities are normalized by δ_j and δ_k, respectively.

3. The previous inequality shows the role of the so-called nonguessing parameter δ_j: a normalization factor to ensure correct comparisons between items and persons. As a matter of fact, equality (4.10) provides us with an interpretation for the parameter δ_j: the difference in answering incorrectly the item standard 1 for two persons $(m = 1, 2)$ must be the same as the difference of these two persons in answering incorrectly any other item provided that the probabilities of answering incorrectly are normalized by the parameter δ_j. In other words, even if an item "invites" to be answered by chance, the differences between persons' characteristics will always be based on their performance in an item that "does not invite" to be answered by chance.

This identification analysis limits the empirical applicability of the 1PL-G model as one can compare the characteristics of stimulus/items and persons only after arbitrarily deciding which stimuli/item will be assumed to have a parameter $c_j = 0$. Assuming that the conclusions will change dramatically if that item is changed seems plausible.

4.4 Identification problems in causal inference

The evaluation of the impact of policy interventions, the effect of a leveling program on students' performance, and the effectiveness of a disease drug are common topics of interest in economic, education, and health-related fields, respectively. In all these contexts, the interest is to recover a treatment effect by comparing the mean outcome difference between sample units under treatment and sample units under the status quo: this corresponds to the so-called ATE; see, for example, Rubin (1974, 1978). Despite the field of application, there is an inherent missing data problem in all treatment effect analyses: each unit in our sample experiences only one of the statuses (treatment/status quo). Thus, "It is a fundamental problem of empirical inference that can be addressed only by making assumptions that relate observed and counterfactual outcomes" (Manski, 2013a, p. 53).

Different approaches have been developed to overcome the unobservability of counterfactual outcomes. In what follows, we revisit the identification of the ATE in the context of an observational study. After that, we will compare this analysis with a partial identification approach.

4.4.1 Point identification of the ATE

Let us begin by explicitly defining the sample space M: it consists of all labels of the sample units (typically persons), and it is, therefore, a finite set. Consequently, the class $\mathcal{M}$ consists of all the subsets of M. Let $\mathcal{T}$ be the set of mutually exclusive treatment indexes. Under these considerations, we define the outcome as follows:

$$Y: M \times \mathcal{T} \to \{0,1\}$$
$$(m,t) \to Y(m,t) \in \{0,1\}$$

where $Y(m,t)$ is the outcome experienced by person m when she/he is exposed to treatment t. Thus, the event $\{m \in M: Y(m,t) = 1\}$ includes *all the persons* in M who have experienced a "positive" outcome when they are exposed to treatment t. The complement of this event represents *all the persons* in M exposed to treatment t and who have experienced a "negative" outcome.

Additionally, we define a random variable (or a function) Z as follows:

$$Z: M \to \mathcal{T}$$
$$m \to Z(m) \in \mathcal{T}$$

where $Z(m)$ indicates the treatment received by person m. Thus, the event $\{m \in M: Z(m) = t\}$ represents all the persons in M exposed to treatment t.

In an observational study, the identified parameters of the statistical model are the following:

(i) The proportion of persons who experienced a "positive" outcome when exposed to treatment t, that is,

$$P\left(\{m \in M: Y(m,t_k) = 1\} \mid \{m \in M: Z(m) = t_j\}\right) \text{ if } t_j = t_k$$

Note that if $t_j \neq t_k$, this probability is not identified because no comparable persons are exposed to treatments t_k and t_j. Therefore, it is impossible to characterize what would have been the outcome of persons exposed to a treatment different from the one they received. This is known in the econometric literature as the common support problem (or assumption) (Lechner, 2008; Blundell & Costa Dias, 2009).

(ii) The proportion of people who received treatment t, namely

$$P(\{m \in M: Z(m) = t\}) \quad \text{for each } t \in \mathcal{T}$$

In what follows, we will analyze the identification problem for two exclusive treatments: innovation, labeled by 1, and status quo, labeled by 0; in this case, $\mathcal{T} = \{0,1\}$. Let us also simplify the notation as follows:

$$\{Y(t) = y\} \doteq \{m \in M: Y(m,t) = y\} \qquad \text{for all } y \in \{0,1\},\ t \in \mathcal{T}$$
$$\{Z = t\} \doteq \{m \in M: Z(m) = t\} \qquad \text{for all } t \in \mathcal{T}$$

From a policymaker's perspective, the interest relies on comparing a "positive" outcome when all persons are exposed to the innovation and the "positive" outcomes when all persons are exposed to the status quo. This is precisely the ATE, namely

$$\mathrm{ATE} = P(Y(1)=1) - P(Y(0)=1) \tag{4.11}$$

To relate the parameters of interest $P(Y(1)=1)$ and $P(Y(0)=1)$ with the identified parameters $P(Y(1)=1 \mid Z=1)$, $P(Y(0)=1 \mid Z=0)$, and $P(Z=1)$, we use the law of total probability (Kolmogorov, 1950):

$$P(Y(1)=1) = P(Y(1)=1 \mid Z=1)P(Z=1) + P(Y(1)=1 \mid Z=0)P(Z=0) \tag{4.12}$$

$$P(Y(0)=1) = P(Y(0)=1 \mid Z=1)P(Z=1) + P(Y(0)=1 \mid Z=0)P(Z=0) \tag{4.13}$$

As we noticed before, $P(Y(1)=1 \mid Z=0)$ and $P(Y(0)=1 \mid Z=1)$ are not identified. Therefore, it is impossible to establish an injection between the parameters of interest $P(Y(1)=1)$ and $P(Y(0)=1)$ and the identified parameters $P(Y(1)=1 \mid Z=1)$, $P(Y(0)=1 \mid Z=0)$, and $P(Z=1)$. In the parlance of causal inference, this is due to the fundamental problem of causal inference:

> It is impossible *to observe* the value of $Y_t(u)$ and $Y_c(u)$ on the same unit and, therefore, it is impossible *to observe* the effect of t on $uY_t(u)tY_c(u)c$. (Holland, 1986, p. 947)

It is typically argued that those parameters of interest can be identified if additional information is collected. Such information is contained in a (vector of) covariate(s) X, namely a random variable (or a function) X: $M \rightarrow \mathcal{X}$, where $\mathcal{X}$ the image space of X, such that $X(m) \in \mathcal{X}$ is associated with each person $m \in M$. Once X is fixed, the parameters of interest are $P(Y(1)=1 \mid X)$, $P(Y(0)=1 \mid X)$, whereas the identified parameters are $P(Y(1)=1 \mid X, Z=1)$, $P(Y(0)=1 \mid X, Z=0)$, and $P(Z=1 \mid X)$. These parameters are related through the law of total probability, namely

$$\begin{aligned} P(Y(1)=1 \mid X) &= P(Y(1)=1 \mid X, Z=1)P(Z=1 \mid X) \\ &\quad + P(Y(1)=1 \mid X, Z=0)P(Z=0 \mid X) \end{aligned} \tag{4.14}$$

$$\begin{aligned} P(Y(0)=1 \mid X) &= P(Y(0)=1 \mid X, Z=1)P(Z=1 \mid X) \\ &\quad + P(Y(0)=1 \mid X, Z=0)P(Z=0 \mid X). \end{aligned} \tag{4.15}$$

The problem of nonidentifiability is analogous to the one mentioned above: $P(Y(1)= 1 \mid X, Z=0)$ and $P(Y(0)=1 \mid X, Z=1)$ are not identified, and consequently, neither are the parameters of interest. The additional information captured in X is used to get identifiability through the so-called strong ignorability conditions (Rosenmbaum & Rubin, 1983), namely

$$P(Y(1)=1 \mid X, Z=0) = P(Y(1)=1 \mid X, Z=1) \tag{4.16}$$

$$P(Y(0)=1 \mid X, Z=1) = P(Y(0)=1 \mid X, Z=0). \tag{4.17}$$

By replacing (4.16)–(4.17) in (4.14)–(4.15), respectively, we get the identifiability of the parameters of interest, namely

$$P(Y(1)=1 \mid X) = P(Y(1)=1 \mid X, Z=1)$$

$$P(Y(0)=1 \mid X) = P(Y(0)=1 \mid X, Z=0)$$

and consequently the ATE as a function of X, namely

$$\mathrm{ATE}(X) \doteq P(Y(1)=1 \mid X, Z=1) - P(Y(0)=1 \mid X)$$

represents a parameter of interest relative to all persons in M with characteristics $X=x$. Once we consider the identification restrictions (4.16) and (4.17), we can point-identify the ATE:

$$ATE(X) = P(Y(1)=1 \mid X, Z=1) - P(Y(0)=1 \mid X, Z=0). \tag{4.18}$$

Note at this point that the restrictions allow us to identify the parameter in those elements in M with characteristics $\{\mathrm{X}=\mathrm{x}\}$ belonging to two mutually exclusive groups, namely $Z^{-1}\{0\}$ and $Z^{-1}\{1\}$, which is a partition of M (here Z^{-1} denotes the preimage of the function Z). It should be remarked that this ATE is a function of the (vector of) covariate(s) X. Using the law of total probability, it is possible to obtain a *marginal* ATE:

$$\mathrm{ATE}(X) = \sum_{x \in \mathcal{X}} \mathrm{ATE}(x) P(X=x)$$

Note that, here, we are restricting all covariates to be discrete.[1]

These results deserve some comments:

1. The strong ignorability conditions (4.16) and (4.17) are *identification restrictions* rather than a component of the model specification. Under this constraint, what a researcher *wants to learn from the data* coincides, for those persons with characteristics $\{m \in M : X(m) = x\}$, with *what can be learned from such data.*
2. Moreover, the ignorability conditions are equivalent to the following conditions:

   $$Y(0) \perp Z \mid X, \quad Y(1) \perp Z \mid X;$$

   for a discussion, see San Martín and González (2022). Using the symmetry property of conditional independence (Florens et al., 1990, chapter 2), it follows that, for $z \in \{0,1\}$,

   $$P(Z=z \mid X, Y(1)) = P(Z=z \mid X), \quad P(Z=z \mid X, Y(0)) = P(Z=z \mid X)$$

   As the reader can recognize, these probabilities are the ones used to estimate the propensity scores and perform the matching procedure. It should be noted that

[1] It is relevant to recall rigorous definition of an absolutely continuous random variable: X is an absolutely continuous random variable if and only if $P(X=x)=0$ for all $x \in \mathcal{X}$. It is (almost) hard to find this type of random variable in concrete applications.

the strong ignorability conditions are well defined once a (vector of) covariate(s) X has been chosen. Consequently, the propensity score procedure is not a procedure to select covariates based on goodness-of-fit indices such as the ones used in confirmatory factor analysis –a widely used method in psychometrics (Brown, 2015): it is only a procedure to perform the matching exercise.

3. The strong ignorability conditions (4.16) and (4.17) do not solve the fundamental problem of causal inference because, according to Holland (1986), such a problem is due to the impossibility to observe *the same statistical unit* exposed to both the innovation $(t=1)$ and the status quo $(t=0)$. For instance, condition (4.16) is an equality between two different mutually exclusive groups of statistical units: those exposed to the innovation, namely $\{m \in M: Z(m)=1\}$ and those exposed to the status quo, namely $\{m \in M: Z(m)=0\}$.

4.4.2 Partial identification of the ATE

Admitting the challenging nature of justifying the strong ignorability condition and the interpretation of the ATE, one might explore the possibility of identifying the parameters of interest using logically weaker identification constraints. Answering this question means moving from point identification of the parameters of interest to partial identification:

> A parameter in a probabilistic model is partially identified if the sampling process and maintained assumptions reveal that the object lies in a set, called the identification region or identified set, that is smaller than the logical range of possibilities but larger than a single point. Sample estimates of partially identified objects generically are set-valued. (Manski, Sanstad, DeCanio, 2021)

In what follows we develop a partial identification approach to partially identify both $P(Y(1)=1 \mid X)$ and $P(Y(0)=1 \mid X)$. We accordingly introduce an identification restriction leading to solving the fundamental problem of causal inference: the conditional probabilities on which these restrictions are based will be conditional on the same statistical units.

Let us explore the impact of the following identification restriction on the identifiability of the parameters of interest:

$$P\ (Y(1)=1 \mid X,\ Z=1) \geq P\ (Y(0)=1 \mid X,\ Z=1) \tag{4.19}$$

$$P\ (Y(1)=1 \mid X,\ Z=0) \geq P\ (Y(0)=1 \mid X,\ Z=0) \tag{4.20}$$

These conditions intend to represent an *optimistic policymaker*. Condition (4.19) implies that individuals exposed to the innovation are less likely to encounter a “positive” outcome under the status quo than when they are genuinely subjected to the intervention. Meanwhile, condition (4.20) means that for those persons exposed to the status quo, it is more probable to experience a “positive” outcome if exposed to the innovation than when exposed to the status quo. In other words, the treatment is better than the status quo for both persons under the innovation and the status quo.

Combining (4.20) and (4.14), we get an interval for all possible values of $P(Y(1)=1 \mid X)$, namely

$$\begin{aligned} &P(Y(1)=1, Z=1 \mid X)+P(Y(0)=1, Z=0 \mid X) \leq \\ &\quad \leq P(Y(1)=1 \mid X) \leq \\ &\quad \leq P(Y(1)=1, Z=1 \mid X)+P(Y(0)=1, Z=0 \mid X)+P(Y(0)=0, Z=0 \mid X) \end{aligned} \tag{4.21}$$

Similarly, combining (4.19) and (4.15), we get an interval for all possible values of $P(Y(0)=1 \mid X)$, namely

$$\begin{aligned} &P(Y(0)=1, Z=0 \mid X) \leq P(Y(0)=1 \mid X) \leq \\ &P(Y(1)=1, Z=1 \mid X)+P(Y(0)=1, Z=0 \mid X) \end{aligned} \tag{4.22}$$

A proof of this result can be found in Appendix B.

To interpret these partial identification intervals, note that

$$\begin{aligned} &P(Y(1), Z=1 \mid X)+P(Y(0), Z=0 \mid X)= \\ &\quad P(\{m \in M\colon Y(m,1)=1, Z(m)=1\} \cup \{m \in M\colon Y(m,0)=1, Z(m)=0\} \mid X) \end{aligned}$$

which corresponds to the proportion of persons who experience a "positive" outcome regardless of whether they are exposed to the intervention or to the status quo. Therefore,

4. Under the optimistic policymaker assumption, the proportion of persons in M who would experience a "positive" outcome if all of them were exposed to the innovation would improve the current proportion of "positive" outcomes. Conversely, the proportion of persons in M who would experience a "positive" outcome if all of them were exposed to the status quo would decrease such proportion.
5. Moreover, the partial identification interval of the ATE(X) is given by

$$\text{ATE}(X) \in [0, P(Y(1)=1, Z=1 \mid X)+P(Y(0)=1, Z=0 \mid X)] \tag{4.23}$$

   that is, it is always positive, and its upper bound corresponds to the current proportion of persons experiencing a "positive" outcome regardless of whether they are exposed to the intervention or the status quo!
6. It should be noted that the point identified $\text{ATE}(X)$ under ignorability conditions (see equality (4.18)) not necessarily is a plausible value of a partially identified $\text{ATE}(X)$ under the optimistic policymaker assumption. As a matter of fact, the point identified $\text{ATE}(X)$ belongs to the partial identification interval (4.23) if and only if

$$\frac{P(Y(1)=1, Z=1 \mid X)}{P(Y(0)=1, Z=0 \mid X)} < \frac{P(Z=1 \mid X)}{P(Z=0 \mid X)}\left(\frac{1}{P(Z=0)}+1\right)$$

Consequently, the conditions of strong ignorability cannot necessarily be interpreted in line with the optimistic policymaker assumption.

7. Indeed, the above conclusions are tautological with the optimistic hypothesis of policymakers. Why is, then, this type of analysis relevant? To answer this, it is necessary to provide an interpretation of the parameters of interest, which requires making explicit the role of the sample space M. Thus, the partial identification (4.21) should be interpreted in the following terms:

   > If all the persons in M (with characteristics $\{X = x\}$) had been exposed to the innovation, then the proportion of those who experienced a "positive" outcome would have been at least equal to the actual proportion of "positive" outcomes, regardless of whether the persons were under the innovation or the status quo. Moreover, this proportion could have increased by a proportion equivalent to the proportion of persons in M (with characteristics $\{X = x\}$) who are under the status quo and who have a "negative" outcome (i.e., $P(Y(0) = 0, Z = 0 \mid X)$ in (4.21)).

   A similar interpretation can be made for the partial identification interval (4.22).
8. This interpretation emphasizes that the evaluation of the policy or program only concerns the population in M, so a policy evaluation should not be confused with a prediction of what might happen if the innovation is implemented. If we want to forecast outcomes for another population $\tilde{M} \neq M$, we would be facing a new identification problem, which is beyond the scope of this chapter.
9. Despite this, it is essential to think about the usefulness of a policy evaluation such as the one above. One possible answer is to consider the concept of *inductive behavior* introduced by Neyman (1938) and developed in Neyman (1950):

   > With many phenomena certain permanencies appear quite stable. This created the habit of regulating our actions in regard to some observed events by referring to the permanencies which at the particular moment seem to be established. This is what we call inductive behaviour. (Neyman, 1950, p. 1)

The interpretation of the partial identification intervals shows what the situation of population M would have been if all of them had been exposed to the innovation or the status quo. But such a situation is a logical consequence of the optimistic policymaker assumption. Consequently, the evaluation of a policy or program aims to persuade the policymaker to *act following this optimistic view*. This is in line with Neyman's inductive behavior concept:

> Nous pouvons *savoir* que la loi mathématique des grands nombres subsiste dans les cas précisés par les conditions des théorèmes qu'on a démontrés. Nous pouvons aussi *savoir* que la loi empirique des grands nombre s'était réaisée dans telles expériences déjà effectueées. Mais nous ne pouvons que *croire* qu'elle continuera à être réalisée dans les expériences futures.
> [. . .]
> Mais se décider à 'affirmer' ne veut pas dire 'savoir' ni même 'croire'. C'est un acte de voloté précédé par quelques expériences et quelqes raisonnements déductifs, tout à fait comme de s'assurer sur la vie, que l'on fait, même si l'on espère vivre longetemps. (Neyman, 1938, p. 352)

In other words, evaluating a policy or a program is intended to modify the policymaker's willingness to act. But let's be clear: it is an evaluation that assumes certain invariance once the population under study is changed.

4.5 How to model self-selection?

Let us continue with the previous discussion on the partial identification of the parameters of interest $P(Y(0)=1 \mid X)$ and $P(Y(1)=1 \mid X)$ in the context of a leveling program of a Chilean public university.

4.5.1 Context

In this Chilean public university, students are selected either by a national admission process (considering high school background and scores from standardized tests) or through an inclusive access program. The Inclusive Access, Equity and Permanence Program (PAIEP, by its Spanish name) is a program developed by the university to support students during their first year at the university. Among other activities, the program considers tutorial classes in both academic and socio-educational topics. Although all students enrolled in the university are invited to participate in this program, the targeted group is the one enrolled through the inclusive access program. Program activities take place throughout the year, although students can stop participating at any time during the year. Furthermore, once the students know their grades for the first semester, the students decide whether to continue in the program during the second semester.

One of the response variables of interest to the program is the grade point average (GPA) at the end of each semester of the first year of university. In addition, the program considers that a student has been intervened if he/she attends at least 10 tutorial sessions per semester.

Once the first semester has ended, students who have participated in the leveling program may choose whether to continue in the program during the second semester. We will now outline how to evaluate the students' decision to continue or not in the program during the second semester. More specifically, for illustrative purposes, we will focus our attention on those students who attended the leveling program during the first semester and obtained a GPA score at least equal to 4.0 (which in Chile is the minimum score to pass a course): we will observe their decision to continue or not in the program during the second semester. Thus, using the notation of the previous section, the labels of these students are gathered in the set M.

4.5.2 Parameters of the problem

For the students in M, we consider the following random variables (always using the notation previously introduced):

- $Z(m) = 1$ if the student $m \in M$ participated in the leveling program during the second semester, and $Z(m) = 1$ if not.
- $Y(m,1) = 1$ if a student $m \in M$ who decides to continue in the leveling program obtains a GPA at least equal to 4.0, and $Y(m,1) = 0$ if he/she obtains a GPA smaller than 4.0.
- $Y(m,0) = 1$ if a student $m \in M$ who decides not to continue in the leveling program obtains a GPA at least equal to 4.0, and $Y(m,0) = 0$ if he/she obtains a GPA smaller than 4.0.
- The covariates X include eventual additional information at the student level.

Thus, the identified parameters are the following:

(a) The proportion of students (with characteristics $\{X = x\}$) who participated in the leveling program on the second semester, namely $P(Z = 1 \mid X)$.

(b) The proportion of students who actually decided to continue in the leveling program and obtained a second semester GPA at least equal to 4.0, namely $P(Y(1) = 1 \mid X, Z = 1)$.

(c) The proportion of students who actually decided not to continue in the leveling program and obtained a second semester GPA at least equal to 4.0, namely $P(Y(0) = 1 \mid X, Z = 0)$.

The parameters of interest are $P(Y(0) = 1 \mid X)$ and $P(Y(1) = 1 \mid X)$, which are unidentified because, as discussed above, $P(Y(1) = 1 \mid X, Z = 0)$ and $P(Y(0) = 1 \mid X, Z = 1)$ are unidentified.

4.5.3 Partial identification analysis

Let us assume that the students' decision to continue or not to continue in the program is "rational" in the sense that if a student decides to continue (relative to the decision to not continue), he/she does so because he/she believes that if he/she does not continue (resp. continues) he/she will have a worse outcome than if he/she continues (resp. does not continue). This assumption can be expressed as follows:

$$P(Y(1) = 1 \mid X, Z = 0) \leq P(Y(0) = 1 \mid X, Z = 0) \tag{4.24}$$

$$P(Y(0) = 1 \mid X, Z = 1) \leq P(Y(1) = 1 \mid X, Z = 1) \tag{4.25}$$

Condition (4.24) implies that among students who actually opt not to continue in the leveling program, if they had chosen to continue, they would have had a lower likelihood of obtaining a GPA $\geq$ 4.0 compared to not continuing in the program. Similarly, condition (4.25) suggests that among students who actually decide to continue in the leveling program, if they had chosen not to continue, they would have had a lower likelihood of achieving a GPA $\geq$ 4.0 compared to continuing in the program.

These identification restrictions imply that

$$\begin{aligned}P(Y(1)=1,Z=1\mid X)\leq P(Y(1)=1\mid X)\leq P(\{&m\in M{:}Y(m,1)=1,Z(m)=1\}\\ &\cup\{m\in M{:}Y(m,0)=1,Z(m)=0\}\mid X)\end{aligned}\tag{4.26}$$

$$\begin{aligned}P(Y(0)=1,Z=0\mid X)\leq P(Y(0)=1\mid X)\leq P(\{&m\in M{:}Y(m,1)=1,Z(m)=1\}\\ &\cup\{m\in M{:}Y(m,0)=1,Z(m)=0\}\mid X)\end{aligned}\tag{4.27}$$

These partial identification intervals deserve the following comments:

- If we are willing to assume that student behavior is “rational” according to assumptions (4.24) and (4.25), then the actual proportion of students obtaining a GPA $\geq$ 4.0 at the end of the second semester will deteriorate if either *all students* in M decide to continue in the program or if they all decide not to continue.
- Considering that the choice is voluntary, what is less bad, to continue or not in the program? One way to respond is to choose the least adverse scenario, which means comparing the lower bounds of the partial identification intervals. That is, we have to compare $P(Y(0)=1,Z=0\mid X)$ and $P(Y(1)=1,Z=1\mid X)$. Assessing whether it is better for students to decide to continue in the program, thus, boils down to comparing the current proportion of students who do not decide to continue and have a GPA $\geq$ 4.0, with the current proportion of students who decide to continue and have a GPA $\geq$ 4.0.

These conclusions show, on the one hand, that believing in a “rational” behavior does not ensure an improvement of the current situation and, on the other hand, they show that the conclusion depends on a given population M and therefore cannot be automatically extrapolated to a different population.

4.5.4 Illustration

We illustrate the previous results with available information regarding the participation of students in the leveling program implemented by a Chilean public university. There are a total of 214 students who attended the leveling program during the first semester and obtained a GPA score at least equal to 4.0. This number represents the cardinality of the sample space M described before.

Because the focus of the PAIEP program is on students selected by the inclusive access program, as an additional characteristic of interest for the students, we define the random variable X – using the notation previously introduced – by $X(m)=1$ if student $m \in M$ was selected by the inclusive access program; $X(m) = 0$ if the student was selected by the national admission process. Thus, for the students selected by the inclusive program, the identified parameters from the available data are:

(i) The proportion of students who participated in the leveling program in the second semester, that is,

$$P(Z=1 \mid X=1) = \frac{153}{201} = 0.7612$$

(ii) The proportion of students who continue in the leveling program and obtained a GPA at least to 4.0 in the second semester, that is,

$$P(Y=1 \mid X=1, Z=1) = \frac{120}{153} = 0.7853$$

(iii) The proportion of students who decided not to continue in the leveling program and obtained a GPA at least to 4.0 in the second semester, that is,

$$P(Y(0)=1 \mid X=1, Z=0) = \frac{33}{48} = 0.6875$$

However, the parameters of interest are $P(Y(1)=1 \mid X=1)$ and $P(Y(0)=1 \mid X=0)$, which represent the proportion of students in M selected by the inclusive program who obtain a GPA at least 4.0 at the end of the second semester, when they decide to continue and not continue in the leveling program, respectively. Under the identified restrictions (4.24) and (4.25), the evaluation of the partial identification intervals (4.26) and (4.27) are as follows:

- $0.5970 \leq P(Y(1)=1 \mid X=1) \leq 0.7612$
- $0.1642 \leq P(Y(0)=1 \mid X=1) \leq 0.7612$

Thus, under a rationality decision assumption, when all students choose to continue or not continue in the leveling program, the proportion of students who obtain a GPA at least 4.0 on the second semester will never be greater than the observed proportion (0.7612). Moreover, considering what was discussed about these intervals in the previous section, and given that the lower bound of $P(Y(1)=1 \mid X=1)$ is greater than the same quantity for $P(Y(0)=1 \mid X=0)$, it would be better for students to continue in the program. We highlight at this point that these results are valid only for the observed students belonging to M and that were selected by the inclusive access program.

4.6 Conclusions and discussion

In this chapter, we have delved into the critical concept of *identifiability* across two fields, namely, econometrics and psychometrics. We have highlighted the significance of identifiability analysis not only in the specification of statistical models but also in conferring statistical meaning upon parameters of interest. A key aspect underscored throughout the discussion is the distinction between "identified parameterization" and "parameters of interest." While identified parameterization pertains to population characteristics and corresponds to functionals of the data-generating process, parameters of interest are substantive issues specific to the data under examination. The crux of the problem lies in establishing a functional injective relationship between identified parameters and parameters of interest. This pursuit of identification is crucial for drawing meaningful and reliable inferences from data. Understanding identifiability is pivotal in ensuring that the parameters we estimate have a valid statistical interpretation and can inform us about the real-world phenomena we seek to study.

As one of the illustrative examples shown in this chapter, the identifiability analysis of the 1PL-G model brings significant clarity not only by resolving the identification issue but also by facilitating the interpretation of crucial parameters of interest, particularly the meaning of guessing, a concept not easy to define. We have seen that the empirical applicability of the 1PL-G model is subject to constraints imposed by the identifiability analysis. The requirement to make arbitrary decisions regarding which stimuli or items will be assigned a parameter $c_j = 0$ presents challenges in comparing the characteristics of stimuli and persons effectively. Consequently, the interpretation and generalizability of the model's outcomes become limited, and we must recognize that alterations to the designated items could significantly influence the resulting conclusions.

As another example, this chapter presents a comprehensive review of the identification conditions for estimating the ATE in observational studies. We offered a formal presentation of the problem, defined the sample space, and identified parameters and parameters of interest. A key point discussed is that the parameters of interest are not identified, leading to limitations in drawing definitive causal inferences. We have argued that the ignorability condition can serve as an identification restriction, enabling the expression of parameters of interest as functions of the identified parameters.

While the ignorability assumption helps to solve the identification problem, it does not fully address the fundamental issue of causal inference, as proposed by Holland in 1986. In response to this challenge, we have presented the concept of partial identification and offered four distinct solutions for causal inference. These solutions draw inspiration from Manski's empirical research approach and Neyman's concept of "behavioral inference", offering promising avenues to overcome the limitations of point identification.

By combining rigorous theoretical analysis with practical approaches, this example contributes valuable insights to the field of observational studies and causal inference. The exploration of partial identification and its integration with established methodologies paves the way for a more nuanced understanding of the ATE estimation in observational settings. Our exposition provides a significant step forward in addressing the identification challenges surrounding the ATE estimation and highlights the importance of considering partial identification methods for advancing causal inference research.

In summary, in this chapter we have shown that the concept of identifiability serves as a cornerstone in statistical analysis, providing a framework for establishing connections between theoretical models and empirical data. By comprehending and addressing the challenges of identification, we can enhance the rigor and validity of our research, ultimately advancing knowledge and understanding across a wide array of disciplines.

Appendix A

The identification analysis of the 1PL-G fixed-effects model with a guessing parameter follows from (4.7). In fact, note that:

$$G\left(\theta_i-\beta_j\right)=\frac{q_{ij}}{\delta_j}\Leftrightarrow\theta_i-\beta_j=G^{-1}\left(\frac{q_{ij}}{\delta_j}\right)\tag{4.28}$$

where $G^{-1}(\cdot)$ represents the inverse function of G. By considering $j=1$ be the standard item (Rasch, 1960), then:

$$\theta_m=G^{-1}\left(\frac{q_{m1}}{\delta_1}\right)+\beta_1$$

which is precisely (4.8). Thus, the meaning of the ability parameter is given by when two different persons indexed by m and l are compared. In fact,

$$\begin{aligned}\theta_m>\theta_l&\Leftrightarrow G^{-1}\left(\frac{q_{m1}}{\delta_1}\right)+\beta_1>G^{-1}\left(\frac{q_{l1}}{\delta_1}\right)+\beta_1\\&\Leftrightarrow q_{m1}<q_{l1}\\&\Leftrightarrow P(X_1(l)=1)<P(X_1(m)=1)\end{aligned}$$

where the second inequality comes from the fact that G is a nonincreasing function and the last one by the definition of q_{mj}. Thus, more ability means greater probability to correctly answer the standard item.

Regarding the parameter β_j, by replacing (4.8) in (4.28), it satisfies:

$$\beta_j=G^{-1}\left(\frac{q_{m1}}{\delta_1}\right)-G^{-1}\left(\frac{q_{m1}}{\delta_1}\right)+\beta_1$$

obtaining equality (4.9). The interpretation of this parameter is obtained by comparing two different items. In fact,

$$\beta_j > \beta_k \Leftrightarrow G^{-1}\left(\frac{q_{m1}}{\delta_1}\right) - G^{-1}\left(\frac{q_{mj}}{\delta_j}\right) > G^{-1}\left(\frac{q_{m1}}{\delta_1}\right) - G^{-1}\left(\frac{q_{mk}}{\delta_k}\right)$$
$$\Leftrightarrow \frac{q_{mj}}{\delta_j} > \frac{q_{mk}}{\delta_k}$$

Regarding to the nonguessing parameter δ_j, when comparing two persons $(m=1,2)$ from (4.9) it follows that:

$$\beta_j - \beta_1 = G^{-1}\left(\frac{q_{11}}{\delta_1}\right) - G^{-1}\left(\frac{q_{1j}}{\delta_j}\right)$$
$$\beta_j - \beta_1 = G^{-1}\left(\frac{q_{21}}{\delta_1}\right) - G^{-1}\left(\frac{q_{2j}}{\delta_j}\right)$$

Given that these results are equal, rearranging terms it holds that:

$$G^{-1}\left(\frac{q_{1j}}{\delta_j}\right) - G^{-1}\left(\frac{q_{2j}}{\delta_j}\right) = G^{-1}\left(\frac{q_{11}}{\delta_1}\right) - G^{-1}\left(\frac{q_{2j}}{\delta_j}\right)$$

recovering equality (4.10). It is important to emphasize that all these results are independent from the function G, the item standard j, and the persons compared.

Appendix B

The identification bounds for the parameters of interest $P(Y(1)=1 \mid X)$ and $P(Y(0)=1 \mid X)$ obtained by using the law of total probability, *Optimistic policymaker* perspective restrictions for the nonidentified probabilities and recognizing that they range in the interval $[0,1]$.

In particular, considering restriction (4.20) in (4.14), the lower bound for the parameter $P(Y(1)=1 \mid X)$ is given by

$$\begin{aligned} P(Y(1)=1 \mid X) &= P(Y(1)=1 \mid X, Z=1)P(Z=1 \mid X) + P(Y(1)=1 \mid X, Z=0)P(Z=0 \mid X) \\ &\geq P(Y(1)=1 \mid X, Z=1)P(Z=1 \mid X) + P(Y(0)=1 \mid X, Z=0)P(Z=0 \mid X) \\ &= P(Y(1)=1, Z=1 \mid X) + P(Y(0)=1 \mid X, Z=1) \end{aligned}$$

The upper is obtained by taking into account that the nonidentified related parameter is a probability, that is, $P(Y(1)=1 \mid X, Z=1) \leq 1$. Then,

$$\begin{aligned}P(Y(1)=1 \mid X) &\leq P(Y(1)=1 \mid X, Z=1)P(Z=1 \mid X)+P(Z=0 \mid X)\\&=P(Y(1)=1, Z=1 \mid X)+P(Z=0 \mid X)\\&=P(Y(1)=1, Z=1 \mid X)+P(Y(0)=1, Z=0 \mid X)+P(Y(0)=0, Z=0 \mid X)\end{aligned}$$

Thus, the lower and upper bound obtained here are precisely the ones shown in (4.21). We emphasize at this point that this interval contains all the possible values for the proportion of person in M who would experience a positive outcome if all of them were exposed to the innovation.

In a similar way, the identification bound for the parameter of interest $P(Y(0)=1 \mid X)$ is derived. As a matter of fact, by replacing restriction (4.19) in (4.15) it holds that:

$$\begin{aligned}P(Y(0)=1 \mid X) &= P(Y(0)=1 \mid X, Z=1)P(Z=1 \mid X)+P(Y(0)=1 \mid X, Z=0)P(Z=0 \mid X)\\&\geq P(Y(1)=1 \mid X, Z=1)P(Z=1 \mid X)+P(Y(0)=1 \mid X, Z=0)P(Z=0 \mid X)\\&=P(Y(1)=1, Z=1 \mid X)+P(Y(0)=1, Z=0 \mid X)\end{aligned}$$

The lower bound for the parameter is attained when the maximum value for the non-identified probability is considered, that is, $P(Y(0)=1, Z=1 \mid X) \geq 0$. Then,

$$\begin{aligned}P(Y(0)=1 \mid X) &\leq P(Y(0)=1 \mid X, Z=0)P(Z=0 \mid X)\\&=P(Y(0)=1, Z=0 \mid X)\end{aligned}$$

Thus, under the last two restrictions mentioned before the identification bound (4.22) is recovered. This interval represents all the possible values compatible with the observables for the proportion of person in M who would experience a positive outcome if all of them were exposed to the status quo.

References

Bahadur, R. R., Stigler, S. M., Wong, W. H. and Xu, D. (2002). R.R. *Bahadur's Lectures on the Theory of Estimation*. Institute of Mathematical Statistics Lecture Notes-Monograph series. Institute of Mathematical Statistics.

Baker, F., & Kim, S. (2004). *Item response theory: Parameter estimating techniques*. New York: Marcel Dekker.

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association*, *72*, 355–366.

Bell, A., Fairbrother, M., & Jones, K. (2019). Fixed and random effects models: Making an informed choice. *Quality & Quantity*, *53*, 1051–1074.

Blundell, R., & Costa Dias, M. (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, *44*, 565–640.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioural Statistics*, *39*, 333–367.

Clarke, P., Crawford, C., Steele, F., & Vignoles, A. (2015). Revisiting fixed-and random-effects models: Some considerations for policy-relevant education research. *Education Economics*, *23*, 259–277.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Engle, R. F., Hendry, D. F., & Richard, J. F. (1983). Exogeneity. *Econometrica*, *51*, 277–304.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, *22*, 309–368.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *17*, 69–78.

Fisher, R. A. (1973). *Statistical methods for research workers*. Hafner Publishing.

Florens, J. P., Marimoutou, V., & Péguin-Feissolle, A. (2007). *Econometric modeling and inference*. Cambridge University Press.

Florens, J. P., Mouchart, M., & Rolin, J. M. (1985). On two definitions of identification. *Statistics*, *16*, 213–218.

Florens, J. P., Mouchart, M., & Rolin, J. M. (1990). *Elements of Bayesian statistics*. Marcel Dekker, Inc.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587–606.

Gourieroux, C., & Monfort, A. (1995). *Statistics and econometric models*. Vol. 1, Cambridge University Press.

Haavelmo, T. (1944). The Probability Approach in Econometrics. *Econometrica*, *12*, iii–115.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Dordrecht: Kluwer Nijhoff Publishing.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Hurwicz, L. (1950). Generalization of the concept of identification. *Statistical Inference in Dynamic Economic Models*, *10*, 245–257.

Itô, K. (1984). *An introduction to probability theory*. Cambridge University Press.

Kolmogorov, A. N. (1950). *Foundations of the theory of probability*. New York: Chelsea Pub. Co.

Koopmans, T. (1949). Identification problems in economic model construction. *Econometrica*, *17*, 125–144.

Koopmans, T. C., & Reiersol, O. (1950). The identification of structural characteristics. *The Annals of Mathematical Statistics*, *21*, 165–181.

LeCam, L., & Schwartz, L. (1960). A necessary and sufficient condition for the existence of consistent estimates. *The Annals of Mathematical Statistics*, *31*, 140–150.

Lechner, M. (2008). A note on the common support problem in applied evaluation studies. *Annales D'économie Et de Statistique*, *91/92*, 217–235.

Longford, N. T. (2012). A revision of school effectiveness analysis. *Journal of Educational and Behavioral Statistics*, *37*, 157–179.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison Wesley.

Manski, C. (2013a). *Public policy in an uncertain world: Analysis and decisions*. Harvard University Press.

Manski, C. (2013b). Diagnostic testing and treatment under ambiguity: Using decision analysis to inform clinical practice. Proceedings of the National Academy of Sciences, 110:2064–2069.

Manski, C., Sanstad, A., & DeCanio, S. (2021). Addressing partial identification in climate modeling and policy analysis. *Proceedings of the National Academy of Sciences*, *11*, e2022886118.

Manski, C. F. (1995). *Identification problems in the social sciences*. Harvard University Press.

McCullagh, P. (2002). What is a statistical model? *The Annals of Statistics*, *30*, 1225–1310.

Mouchart, M., & Oulhaj, A. (2006). The role of the exogenous randomness in the identification of conditional models. *Metron*, *64*, 253–271.

Neyman, J. (1938). L'estimation statistique traitée comme un problème classique de probabilité'. *Actualités Scientifiques Et Industrielles*, *739*, 25–57.

Neyman, J. (1950). *First course in probability and statistics*. Rinehart and Winston: Holt, Inc.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. The Danish Institute for Educational Research.

Rasch, G. (1966). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.). *Readings in mathematical social sciences* (pp. 89–107). MIT Press.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*, 34–58.

San Martín, E., del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, *30*, 183–203.

San Martín, E., & González, J. (2022). A critical view on the NEAT equating design: Statistical modeling and identifiability problems. *Journal of Educational and Behavioural Statistics*, *47*, 406–437.

San Martín, E., González, J., & Tuerlinckx, F. (2009). Identified parameters, parameters of interest and their relationships. *Measurement: Interdisciplinary Research & Perspective*, *7*, 97–105.

Van der Linden, W., & Hambleton, R. (1997). *Handbook of modern item response theory*. New York: Springer.

Weitzman, R. A. (1996). The Rasch model plus guessing. *Educational and Psychological Measurement*, *56*, 779–90.

Jeanette Melin

5 Is validity a straightforward concept to be used in measurements in the human and social sciences?

Abstract: Validity in measurements in human and social sciences is commonly referred to as "measuring what one intends to measure," and with a good fit of item parameters – somewhat simplified – it is considered to ensure validity when measuring latent traits in persons. Despite new thinking and trends about validity and positioning validity in measurement theory and practice, today's use of validity can mostly be traced back to the classical test theory (with no compensation for ordinality, no proper separation of latent traits for persons and items, nor a defined measurement system). Consequently, when positioning models, measurement, and metrology to extend the SI, there is a need to critically review the concept of validity. A fundamental mistake is that, too often, a proper distinction is not made between the latent trait itself and the latent trait as measured. In the human and social sciences, where there are yet to be any established units, measurement-related validity should ideally not precede validity in the latent trait itself. Notably, the concept of validity has so far not been included in the International Vocabulary of Metrology (JCGM, 2012), although validation processes (entry 2.45) have been included. This is reasonably due to the centuries of work contributing to a solid consensus about the quantities in themselves. However, given the urgent needs of society for new knowledge about the world to make well-informed decisions about measurements of latent traits, we do not have centuries to first reach consensus about measurement validity. Neither was this done with the existing SI, which has been an iterative work, defining the quantities and measurement processes. Therefore, in a time where the possibilities for new units to extend the SI are being explored, an iterative and cross-disciplinary effort is needed. Thus, this chapter reviews and discusses validity and its related aspects. Finally, the chapter concludes with a proposed call for action to include a nuanced view of validity when extending models, measurement, and metrology of the SI to include measurement in the human and social sciences.

Keywords: validity, construct theory, construct specification equations

Acknowledgments: This chapter is based on learnings and outcomes in several projects during the past years, including the RISE internal platform for Categorical Based Measures and strategic initiative Measurement for Sustainable Transformation. The author would like to give a particular thanks to Leslie Pendrill and Marie-Louise Möllerberg for fruitful discussions on the topic.

Jeanette Melin, Division of Safety and Transport, Department of Measurement Science and Technology Unit, RISE Research Institutes of Sweden, Gothenburg, Sweden; Department of Leadership, Demand and Control, Swedish Defence University, Karlstad, Sweden

https://doi.org/10.1515/9783111036496-005

5.1 Why it is important to care about validity in measurements

Measuring is never an end itself; rather, it is a way of gaining knowledge about the world to make well-informed decisions. To quote Fisher (1994), *because we intend to use our measures to inform decisions that affect people's lives, we are ethically bound to be sure that the numbers represent more or less of the construct in question.* This is a general matter and not unique to measurements in the human and social sciences. However, as will be shown in this chapter, measurements in the human and social sciences face other challenges than measurements in physics, where validity is a critical component. Validity is of course important in both areas, but today for measurements in the human and social sciences, we face different challenges than in physics, and where validity is central is a critical component.

To give some examples of decisions based on measurements of latent traits in the human and social sciences, in health care, it could be questions about giving a diagnosis or prescribing treatment, setting school grades or providing support for learners with special needs, and in recruiting for work, it could be used in personnel selection (Newton & Shaw 2014). In all these cases, and many more, if the measurement does not validly capture the latent trait of interest, such decisions cannot be made validly and reliably. In health care, for instance, some patients as a consequence might be denied treatment while others who do not need it will get it, to name but one example of the serious impact of a lack of validity.

In decision-making based on measurements, it is not only validity that is important but also reliability. Figure 5.1 provides a classic picture of four dichotomous cases of measurement results to be either valid or not, and reliable or not.[1] A valid and reliable measurement is, of course, optimal in any decision-making; on the contrary, a measurement that is neither valid nor reliable is most often meaningless. In between, there is a trade-off between validity and reliability (Clifton, 2020); validity may increase with a decrease in reliability, and validity may decrease with an increase in reliability. However, the usefulness of a reliable but not valid measurement is questionable; we only know that we are measuring something well, but we do not know what we are measuring. Thus, in line with most psychometricians, we argue that validity is paramount, and reliability is contingent upon validity in measurements in the human and social sciences (Johansson et al., 2023). Furthermore, it should be emphasized that neither rigorous research design, advanced statistics, nor large samples (Flake & Fried, 2020) can make an invalid measurement valid afterward. Therefore, in the light of positioning models, measurement, and metrology to extend the SI, there is a need to critically review the concept of validity in accord with the purpose of meas-

[1] Note that this cross table also applies to the latent trait and the decision based on the latent trait as measured, which will be addressed in the forthcoming sections.

urements as a way of gaining knowledge about the world to make well-informed decisions, even those based on latent traits.

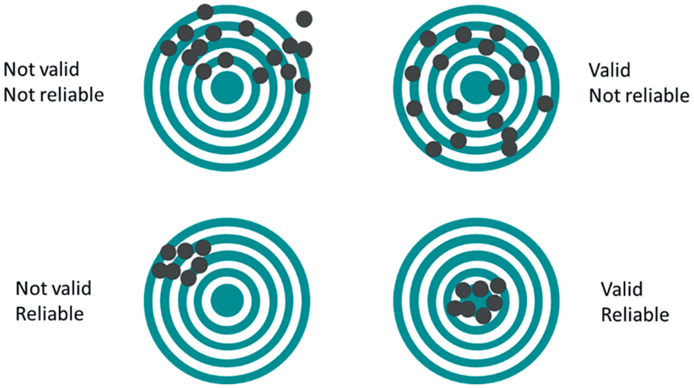


Figure 5.1: Four cases where a measurement can be either valid or not, as well as reliable or not. Dots that are closer to the middle indicate better validity, and more consistency of the dots indicates better reliability.

5.2 Latent traits and measurements of latent traits

For the traditional physical quantities and SI units (length, mass, time, etc.), for centuries, there have been internationally established agreements and definitions of the quantities themselves and procedures for measuring the quantities. However, for latent traits, there are neither such agreements about the latent traits themselves nor for the procedures of measuring the latent traits. Too often, a fundamental mistake is made when aiming to measure a specific latent trait before even understanding the existence of the latent trait itself. Therefore, we start this section by addressing latent traits themselves, followed by how to measure them.

Importantly, this chapter will not provide a discussion either on *if* latent traits exists or *if* latent traits can be measured, and such discussions, summaries, and reflections can be found elsewhere (cf. Finkelstein, 2003; Maul, Torres Irribarra & Wilson, 2016; Slaney, 2017; Michell, 2021; Mari, Wilson & Maul, 2022). Rather, when positioning models, measurement, and metrology to extend the SI into measurements in health and social sciences, our point of departure is that latent traits exist, and thus, they can be measured.

It should be noticed that this distinction between a latent trait and a latent trait *as measured* is, of course, equally important for quantities themselves and quantities as measured (Pendrill, 2019), but as will be shown later in this chapter, it has too often been forgotten, which in turn contributes to the inconsistent use of validity in measurements in human and social sciences. It is further worth to note that there is a hierarchical differentiation of quantity-related concepts and relations, as can be read in more detail in the accompanying chapter by Pendrill (2024) in this monograph. Building on the work of Dybkaer (2010) and others concerning quantities in general, but equally applicable for latent traits, concepts range from the superordinate *kind of quantities* to more specific terms, such as *quantities*, *entity quantities*, and *instantiations of an entity quantity*. Quantities as measured, as well as latent traits as measured, only have a full associative relation to instantiations of an entity quantity specifying quantity *X* itself, for entity *Y* at time *Z* to quantity *X* as measured, and for entity *Y* at time *Z* (Pendrill, 2019). However, for the readability in this chapter, we will use the shorter form: latent trait itself and latent trait as measured.

5.2.1 Latent traits

Latent traits are "hidden" variables, typically proposed to be within a person[2] (cf. Tesio, 2003; Battisti, Nicolini & Salini, 2010; Tesio et al., 2023) and often the main interest of the end-users. However, as will be further emphasized below, latent traits are also attributed to tasks. Examples of latent traits related to decisions in Section 5.1 could be a specific ability of a patient important for setting a diagnosis or prescribing treatment, the learner's math ability to provide grades or support for special needs, or person's attributes important in personnel selection. The basic idea is that the latent trait is unobserved but can be accessed via observed (manifest) variables conditionalizing the latent trait (Borsboom, Mellenbergh, & van Heerden, 2003). Thus, whether one observes manifest variables conditionalizing the latent trait or not, the latent trait itself can exist. Similarly, a person's mass or temperature (i.e., the quantity itself) exists independently of whether it is being measured or not, that is, at least for physical quantities there is a strong objectivity.[3]

While latent traits belonging to persons are very much what end-users are interested in the human and social sciences, nevertheless, for researchers, metrologists, psychometricians, and so on, there is another class of latent trait coupled to the latent trait of the person (Pendrill, 2014), namely latent traits attributed to tasks, which are

[2] An "agent" is the entity term superordinated to persons, organizations, cities, and so on, while patients, learners, and recruiters are subordinated to persons. The utmost work in the human and social sciences is related to persons (including subordinated groups); therefore, we use that term consistently throughout the chapter.

[3] Traditionally, a corresponding, albeit weak, objectivity is taken to apply in the human and social sciences (Pendrill 2019); a contrary position holds that the objectivity obtained both in (a) Bohm's (1952; Bohm & Hiley, 1984, 1989; Bohm, et al., 1987) ontological interpretation of quantum mechanics (Esfeld, et al., 2014; Goldstein, 1998a/b, Matarese, 2023), and in (b) Prigogine's (1971, 1976, 1978) entropy-driven theory of dissipative structures offers a potential philosophical unification of the sciences (Bohm, 2005; Bohm & Hiley, 2006; Prigogine & Stengers, 2018) exemplified by the perspective on measurement taken here (Fisher, 1988, 2024).

of significance when continuing to measurements of latent traits of persons. This is, unfortunately, unknown or neglected by many. A reason why the latent trait attributed to tasks has been less emphasized in the literature could be associated with the analogies often made between questionnaires and engineering instruments (such as thermometers). Interestingly, even though they are estimated from the margins of the same conjointly ordered data matrix, it is rarely seen that the latent trait for the items is referred to as "measured" to the same extent as measures of the latent trait for the person (see further discussion at the end of Section 5.2.2).

In fact, latent traits in themselves have nothing to do with the measurement processes. One way of illustrating this is presented in Figure 5.2, where a latent trait of a person is represented by θ_i and a latent trait of a task by δ_j; both of them exist independently of each other and can be defined as $g(\chi_i, \varepsilon_i, \beta)$ and $f(\upsilon_j, \epsilon_j, \gamma)$, respectively. Thus, both latent traits have their unique sets of explanatory variables and unexplained parts. Nevertheless, how we understand and define latent traits will, of course, have implications for the measurement process, which will be addressed in the following sections of this chapter.

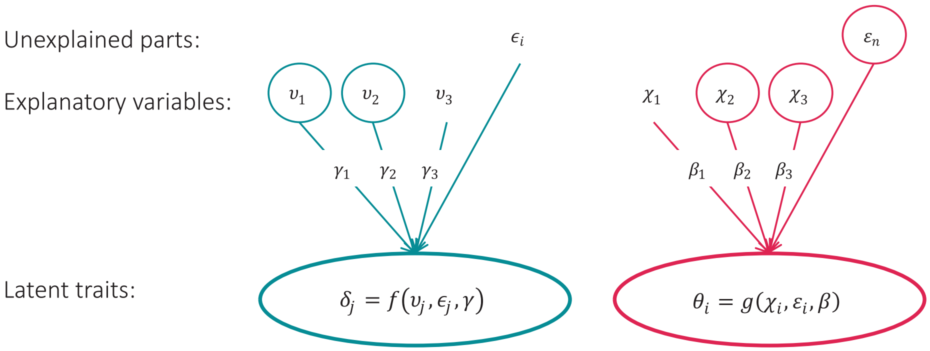


Figure 5.2: Illustration and notations of two coupled latent traits, for example, for tasks δ_j and persons θ_i, and how they are presented as a function of both explanatory variables and unexplained parts.

5.2.2 Measurement of latent traits

When positioning models, measurement, and metrology to extend the SI to include latent traits, a wide definition of measurement is an important starting point (Finkelstein, 2003). Measurement can be defined as an *empirical operational procedure which assigns numbers to members of a class of entities, in such a way as to describe them; by which is meant that relations between these numbers correspond to empirical relations between the entities to which they are assigned* (Finkelstein 1975). Therefore, a critical first step toward measuring latent traits is to observe manifest tasks representing the latent trait of interest to determine how much or how little a person has of the latent trait of interest. In the simplest form, when observing manifest tasks, a person can either pass or fail, typically classified as one [1] if the test person passes or zero [0] if the test person fails. However, such classifications have no numerical meaning and only serve to indicate ordered categories (for nominal data, the categories are not ordered). Despite this well-known fact about ordinal data, counting raw scores or calculating the probability of success in a test as a measure of a test person are, unfortunately, still practiced in many fields but lack metrological quality assurance.

The basic idea of observing manifest tasks representing the latent trait of interest is very similar when advancing the methods to ensure measurement quality. In fact, as has been noted for some decades (Andrich, 1978, 1988, p. 43; Engelhard, 2012; Linacre, 1995, 2000a/b; Wright, 1997), multiple independent developments (Bradley & Terry, 1952; Luce, 1959; Luce & Tukey, 1964; Peirce, 1878; Rasch, 1960; Thurstone, 1928; Zermelo, 1929) show that ordinal observations can be restituted into interval measurements via models defining unit quantities that retain their properties independent of the questions asked and persons responding to within a fit-for-purpose degree of uncertainty. Thus, there are two critical phases for providing measurements of latent traits:

- the observation phase, that is, when data is collected, for instance with a questionnaire, observation protocol, or test from a person or a group of persons, and
- the restitution phase, that is, when separating the probability of success from the observed data into separate measures of the two latent traits (attributed to persons and tasks).

When considering the basic assumption of measuring latent traits – that a person who has more of the latent trait will be more likely to score higher on a difficult item (i.e., manifest task) than a person who has less of the latent trait, and conversely, it is more likely that more persons score high on an easy item – the importance of quantifying both latent traits and their relationship might become clearer. This relation between the two coupled attributes is given by the formula (Rasch 1960; Wright & Stone, 1979):

$$P\left(\pi_{ij} = 1 | \theta_i, \delta_j\right) = \frac{e^{\left(\theta_i - \delta_j\right)}}{1 + e^{\left(\theta_i - \delta_j\right)}} \tag{5.1}$$

where the probability π of a response scored 1 from person i in relation to task j is a function of the difference between θ_i, the latent trait attributed to the person, and δ_j, the latent trait attributed to the tasks. Rasch's (1960, pp. 110–115) formulation of this model originated in an analogy from Maxwell's treatment of Newton's second law of motion, but most early applications were in the educational sciences, where the common association of latent traits attributed to persons are typically referred to as abilities and latent traits attributed to tasks are typically referred to as difficulties. In this chapter, when we provide examples, for simplicity, we will use person's ability and task's difficulty, but the thinking, of course, can also be extended to other latent traits.

Figure 5.3 links the latent traits themselves (Figure 5.2) via the observation phase and the measurand restitution (eq. (5.1)) to the latent traits as measured (for items $\delta_{j,m}$ and persons $\theta_{i,m}$). At the top of the figure, we have two latent traits that need to be "coupled," and the observed response depends on both the latent trait attributed to the person and the latent trait attributed to the task. As we often intend to measure the abilities of persons, we can also refer to the observation phase where the observed response will depend on the person's ability and the item's difficulty. In the next step, with measurand restitution – here done by estimating the parameters in eq. (5.1) – separate measurements of the coupled latent traits can be obtained. Thus, we are using the manifest tasks to provide measurements of the latent trait of persons, and we are using persons to provide measurements of the latent trait of the tasks.

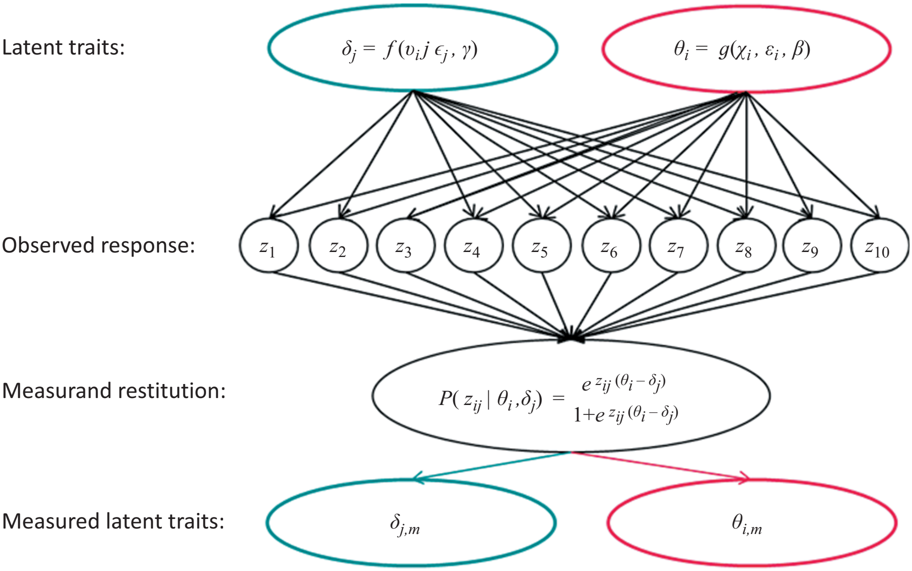

Latent traits:
$\delta_j = f(\upsilon_i j\, \epsilon_j, \gamma)$
$\theta_i = g(\chi_i, \varepsilon_i, \beta)$
Observed response:
z_1 z_2 z_3 z_4 z_5 z_6 z_7 z_8 z_9 z_{10}
Measurand restitution:
$P(z_{ij} \mid \theta_i, \delta_j) = \frac{e^{z_{ij}(\theta_i - \delta_j)}}{1 + e^{z_{ij}(\theta_i - \delta_j)}}$
Measured latent traits:
$\delta_{j,m}$
$\theta_{i,m}$

Figure 5.3: Illustration and notations for how the latent traits themselves (Figure 5.2) are coupled into the observation phase and through the restitution phase provide separate measures of latent traits, for example, tasks $\delta_{j,m}$ and persons $\theta_{i,m}$.

The observed response does not, however, depend only on the latent trait of the person and the latent trait of the task. There are, in fact, additionally other components from the measurement process that are not yet fully compensated for in the model shown in eq. (5.1). For example, Figure 5.4 shows a more complete picture for latent traits of the measurement process (Pendrill, 2019; Pendrill, 2014; Bentley, 2005; Pendrill, 2023), where measurement information is transmitted from the measurement *object,* via an *instrument* to an *operator* in the observation phase, which both the *environment* and the *measurement method* can influence.

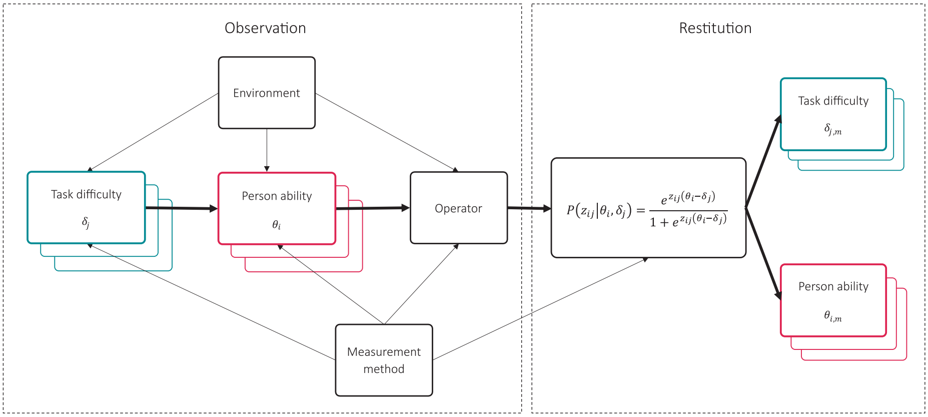


Figure 5.4: An illustration of the measurement system for latent traits linking the observation phase with the restitution phase. Tasks, for example, questionnaire items, provide stimuli due to their difficulty to the test persons who respond to each item, where the response depends on both the difficulty of the task and the ability of the person (i.e., the latent trait themselves, for items δ_j and person θ_i), which in turn can be restituted with the model shown in eq. (5.1) into measurements of tasks $\delta_{j,m}$ and persons $\theta_{i,m}$.

This view of the measurement system corresponds more directly with the traditional approach in engineering science and technology than with typical arguments in human and social sciences measurements. Much is to be gained by adopting this approach. Specifically, in measurement engineering (Bentley, 2005), an instrument converts an input (such as from a stimulus from the measurement object) into an output response, while the measurement object has no input but only produces an output (which acts as a stimulus input to the instrument), for example from weighing, where an object has a mass that stimulates the instrument (scales) to respond with an indication of the mass. Similarly, in both traditional and "psychometric" measurement systems, the measurement object (weight or task) is a natural first choice of metrological standard – with its robustness and simplicity – in preference to the relatively sensitive and complex instrument, with more sensitivity to the environment, context, and method (Pendrill, 2021; Melin, 2021). Thus, Pendrill (2018) has argued that: *drawing simple analogies between "instruments" in the social sciences questionnaires, ability tests, etc. – and engineering instruments such as thermometers does not go far enough*. As will be shown later in this chapter, a complete picture of the measurement process and the measurement system will have implications for using the concept of validity, which is significant when positioning models, measurement, and metrology to extend the SI.

Notably, in contrast to measurements in physics, calibration and the measurement itself are often done simultaneously for measurements in human and social sciences. For example, while arguing that a set of items, that is, an item-bank, is analogous to a calibrated set of weights to ensure metrological comparability when measuring person's ability (Pendrill, 2018), previously existing measurements of task's difficulty are not always being used for measuring person’s ability in a new cohort. Nevertheless, with the model shown in eq. (5.1) (Rasch 1960), measurements of person’s ability are easily restituted based on the raw scores from the observation phase based on previously existing measurements of task’s difficulty. In turn, this will enable comparability beyond the present cohort of persons. Another way, even more accessible, to achieve measurements of person’s ability is, thanks to conversion tables, again where raw scores from the observation phase are being used and converted to measures in the same way that meters can be converted to inches (Melin et al., 2023a).

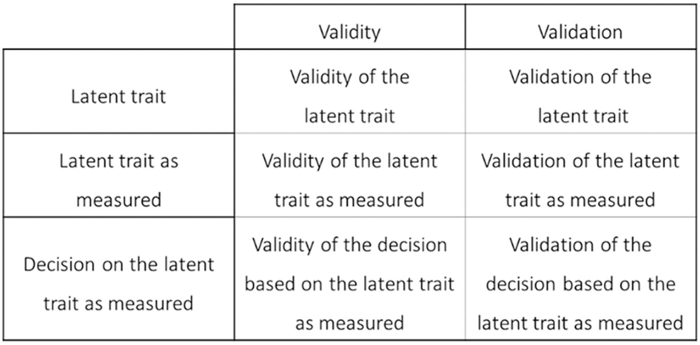
5.3 Validity and latent traits

History shows that validity concepts in measurements in human and social sciences has undergone a “metamorphosis” (Geisinger 1992). Although validity in measurements in human and social sciences is commonly referred to as *whether a test measures what it purports to measure* (Kelley 1927), it is only sometimes reflected in practice when choosing theories and methods. Many others have made good summaries of the evolvement of validity as a concept in the human and social sciences (cf. Messick 1989a; Newton & Shaw, 2014; Borsboom, 2005; Slaney, 2017; Kane, 2016), and such summaries go beyond the scope of this chapter. However, we will instead pick up some of the key contributions to today’s somewhat fragmented use of validity and will, at the end of this section, return to and review the statement by Borsboom et al. (2004), claiming that *validity is not complex, faceted, or dependent on nomological networks and social consequences of testing.*

5.3.1 Validity and validation

The first very fundamental differentiation is between validity and validation: validity is about ontology, and validation is about epistemology (Borsboom, Mellenbergh, & van Heerden, 2004). First, we argue for the need to consider validity aspects related to both the latent traits themselves (Section 5.2.1) and the latent traits as measured (Section 5.2.2), that is, the trueness of both the existence of the latent trait and of the measurement results. Furthermore, Wolf et al. (2019) summarized the contemporary validity literature as saying that *validity is not an inherent feature of a survey (or other instruments) but rather a characteristic of the survey concerning a particular use* [. . .] *as a consequence, validation is necessarily fit-for-purpose, such that different forms of argumentation and evidence may be necessary depending on the design and intended purposes of the survey.* This gives us three potential validity claims: the validity of the latent trait, the validity of the latent trait as measured, and the validity of the decision based on the latent trait as measured.

The literature has been moving away from the concept of validity and emphasizing instead methods for validation, which are – at least theoretically – also applicable to all three validity claims. Figure 5.5 summarizes those distinctions for any latent trait. However, below, it will be put in the context of the two coupled attributes: a latent trait attributed to persons and a latent trait attributed to tasks. In addition, it should be noted that there is a further question: Is the validation valid? That is, one must distinguish between two kinds of decisions: the validity of the decision on the latent trait as measured and the validity in the claim about the validation of the latent trait.



	Validity	Validation
Latent trait	Validity of the latent trait	Validation of the latent trait
Latent trait as measured	Validity of the latent trait as measured	Validation of the latent trait as measured
Decision on the latent trait as measured	Validity of the decision based on the latent trait as measured	Validation of the decision based on the latent trait as measured

Figure 5.5: The distinction between validity aspects and validation for the latent trait, the latent trait as measured, and the decision based on the latent trait as measured.

When extending models, measurement, and metrology of the SI into measurements which also cover the human and social sciences, of course, one needs to consider the International Vocabulary of Metrology (VIM) (JCGM, 200:2012). Notably, validity is not yet included in the VIM, while validation (entry 2.45) is. This reasonably is due to the centuries of work to reach a consensus on the physical quantities[4] in themselves. However, an important note is that validation is defined as: *verification, where the specified requirements are adequate for intended use*, reflecting validation of the measurement process rather than validation of the quantities themselves or quantities as measured.

To summarize this section, validity and validation are distinct concepts that should not be mixed. One must be careful when making claims about measurement-related validity and decision-related validity before the validity in the latent trait itself has been ensured. In a time where the possibilities to extend the SI for new units are being explored, it is, however, important to stay open for an iterative and cross-disciplinary effort to advance both the validity of the latent traits themselves and the measured la-

[4] Since 1968, within the SI units, there are not only physical quantities but also mol for the amount of substance.

tent traits as well as developing methods for validation of latent traits themselves and the measured latent traits. This will be further discussed in Section 5.4.

5.3.2 The many facets of validity

A decade ago, Newton and Shaw (2014) published a book about validity in educational and psychological testing, including a list of 151 (!) different kinds of validity. Based on decades of research, they summarized three different claims related to validity:

1. Validity as a **measurement claim**: It is possible to measure a latent trait accurately using a measure of the latent trait.
2. Validity as a **measurement and decision-making claim**: It is possible to make accurate decisions on the basis of measurements of the latent trait.
3. Validity as a **concept spanning measurement, decision-making, and broader impacts and side-effects**: It is acceptable to implement a measurement policy.

While none of these actually address the validity of the latent trait itself, that is, if the latent trait exists or not, the first claim relates very much to the original statement of validity – *whether a test measures what it purports to measure* (Kelley 1927). This is also related to the "middle validity claim" in Figure 5.5 (i.e., the validity of the latent trait as measured). The second and third validity claims are reasonably a response to the significance of being able to justify interpretations and actions concerning social and ethical consequences of test use (Messick 1989a, 1989b) and the separation of different kinds of validity (Joint Committee on the Standards for Educational and Psychological Testing, 2014; Cronbach & Meehl 1955). Tracing back to the mid-nineteenth century, three types of validity dominated, namely content, construct, and criterion validity. Traditionally, both content and construct validity relate to how a set of test items can be used to measure a person's latent trait of interest validly; content validity is typically referred to if the set of test items reflects the important components related to a given person's latent trait and construct validity on the psychometric properties of the used set of items. On the other hand, criterion validity is more related to how measurement values of the person's latent trait can either be compared with results from similar measurements (also known as concurrent validity) or predict an outcome at a later time (also known as predictive validity). Those three types of validity have different significance in different contexts, where content validity has a particular role in achievement tests, construct validity for personality tests, and criterion validity for an aptitude test (Newton & Shaw, 2014). We acknowledge this tradition and understand that different aspects may have different importance for the end-user, but simultaneously believe that this confuses the use of the validity term.

The use of criterion-related and, in particular, predictive validity has been and continues to be dominating in personal selection. For this purpose, the persons' measurement values of the latent trait are viewed as a *sign or signal of future performance* *and rely on evidence that individuals with higher predictor scores* [i.e., measurement value] *subsequently perform better* (Van Iddekinge, Lievens, & Sackett, 2023). Thus, the main focus is on the measurement value and its relation to the future rather than what the measurement value stands for. Roughly speaking, if measurement values from persons based on a set of items can predict future outcomes well, then the prediction is more important than the latent trait itself and the latent trait as measured. Here, greater "allowances" to focusing on reliability at the expense of validity are often accepted (Clifton, 2020).

Since the 1950s, the American Psychological Association, the American Educational Research Association, and the National Council on Measurements in Education have been leading actors in the field of validity, including publishing of the *Standards for Educational and Psychological testing*. Initially, focus was on the three parts of validity (i.e., content, construct, and criterion validity), but in the third (and fourth) edition, there has been a shift toward considering validity as multidimensional and complex, requiring a wide and diverse body of evidence (Goodwin & Leech, 2003). The *Standards* comprise the following validity-related concepts (Joint Committee on the Standards for Educational and Psychological Testing, 2014):

- Evidence-based **test content** refers to the set of items that represents the domain it proposes to measure (similar to content validity)
- Evidence-based **response processes** are the extent to which different types of respondents' responses fit the defined construct (similar to construct validity)
- Evidence-based **internal structure** is about how the components match the defined construct (similar to construct validity)
- Evidence-based **relations to other variables** reflect expected relations based on the theory of the construct being assessed (similar to criterion validity)
- Evidence for **validity and consequences of testing** includes both anticipated and unanticipated consequences of the measurement.

Building on McAllister's (2008) claim that probabilistic conjoint measurement offers *a statistical model for validating assessment tools that are particularly suited to quantifying human performances on assessment items*, Mui Lim et al. (2009) proposed examples of validation activities and validation linked to the types of validity in the *Standards* (Joint Committee on the Standards for Educational and Psychological Testing, 2014). While we would, in line with Pendrill (2014), stress that identified measurement models' (San Martin & Rolin, 2013) testing for conjoint additivity (Newby & Bunderson, 2009), parameter separation, and specific objectivity (Rasch, 1966) are *not simply a mathematical or statistical approach but instead a specifically metrological approach to human-based measurement*, the proposal of how Mui Lim et al. (2009) suggests validation activities are very welcomed in relation to the view of validity provided by the *Standards*. On the contrary, in view of the weakness in the *Standards* of not thoroughly addressing validity in the latent trait itself, such validation activities can be carefully implemented, provided one has firstly ensured the validity of the latent trait itself.

Furthermore, it has been proposed that the statistical sufficiency principles of measurement modeling (i.e., that the observed score should capture all available information in the data; Andersen, 1999; Andrich, 2010) are related to Messick's construct validity issues (Baghaei, 2008), namely, *that nothing important be left out* (Messick 1996; Messick 1994). Here, Baghaei (2008) argues for assessing model fit statistics to indicate possible construct-irrelevant variance and assessing the conjoint item-person histogram to assess construct underrepresentation. He further links some different types of validity to typical evaluations as to whether the items address the intended latent trait, the item difficulty hierarchy makes sense as an expression of the construct, and measurement values correlate well with measurements estimated from other sets of items probing the same latent variable. On the other hand, questions as to whether the person's ability hierarchy makes sense indicate that this approach to predictive validity may be less related to the common practice of comparing measurements to an outcome estimated via other means at a later time.

This section has provided a short summary of the many facets of validity, including some work specifically related to probabilistic conjoint measurement (as it is key in the measurement restitution process). Work, so far, has been dominated by the latent trait as measured for persons. This can reasonably be explained by end-user interest in making decisions about the persons based on measurement. It is also likely that the classical test theory – that is, where no proper separation is made between the latent traits of persons and items are being made – has impacted the setting of the terminology and use of it. Nevertheless, we agree with Stenner (2014) that *validity should be equally applicable to both latent traits* in measurements in the human and social sciences.

5.3.3 Revisiting the bull's eye target

It appears that Figure 5.1 – the classic bull's eye target for illustrating validity and reliability – could be revisited by asking what target is aimed for the closeness to the latent trait itself or the closeness to the measurement of the latent trait based on a reference method? Figure 5.6 highlights four possible different targets for the bull's eye. Obviously, the closeness to the latent trait itself relates to the overall aim of measuring, that is, a way of gaining knowledge about the world to make well-informed decisions. At the same time, comparing a measured value of a latent trait with a value of the latent trait itself is impossible because the value of the latent trait itself is not accessible. Despite that, it is likely that much of the literature refers to the middle point as the latent trait itself when not properly distinguishing between the latent trait itself and the latent trait as measured.

In an illustrative case (Figure 5.7), imagine analyses of two questionnaires claiming to measure the same latent trait of persons but with different sets of items. Suppose that both fit the measurement model well (e.g., no significantly misfitting items, no differential item functioning, no local dependency beyond the modeled stochasticity of the

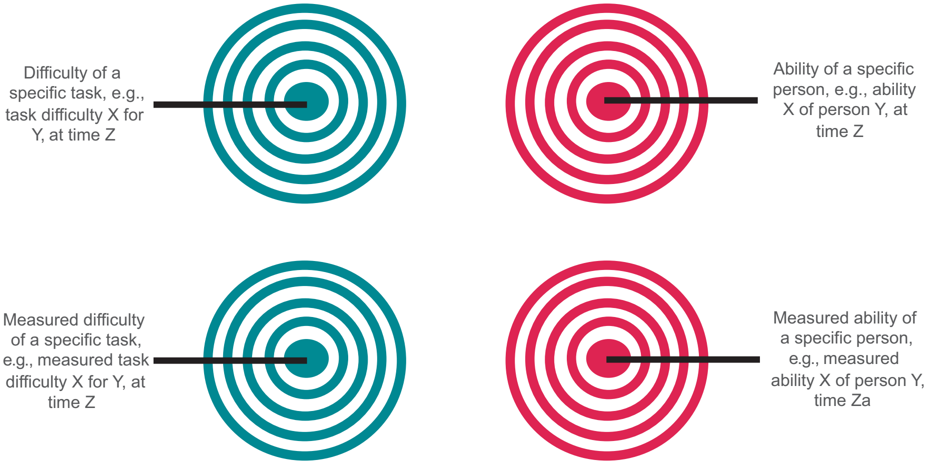

Difficulty of a specific task, e.g., task difficulty X for Y, at time Z
Ability of a specific person, e.g., ability X of person Y, at time Z
Measured difficulty of a specific task, e.g., measured task difficulty X for Y, at time Z
Measured ability of a specific person, e.g., measured ability X of person Y, time Za

Figure 5.6: Four cases with different possible targets in the bull's eye. To the left, targets are related to task's difficulty and to the right targets are related to person's ability. The upper boards are related to the latent trait themselves, and the lower boards are related to the latent traits as measured.

data, unidimensionality usefully approximated to within the tolerance limits of the application, and response categories that work as intended (Johansson et al., 2023)), but when correlating to another variable, the association is different. This raises questions such as how to make a valid inference about the association between y and the latent trait and which one is the "optimal" way of gaining knowledge about the world to make well-informed decisions? In fact, similar results have empirically been shown by Maul (2017), for instance, by including two set of items intended to measure growth mindsets with the notable exception that the key noun in the sentence ("intelligence") had been replaced with a nonsense word ("gavagai") in one of the item sets. Analyses of both sets of items, however, fitted models well. To quote Maul (2017), *it would seem difficult to take seriously the claim that any of these sets of items constituted a valid measure of a psychological attribute, and if such a claim were made, one might reasonably expect any quality-control procedure worthy of the name to provide an unequivocal rejection*. Thus, in cases where interpretation of the bull's eye becomes a measurement issue, it cannot be separated from a qualitative understanding of the latent trait itself.

5.3.4 Designing measurements of latent traits

As shown in Figure 5.2, a latent trait of a person, θ_i, and a latent trait of a task, δ_j, can exist independently of each other, while they show a special relationship when measuring latent traits (eq. (5.1)). To respond to the most common end-user need – measurement values of a specific latent trait attributed to persons – it is natural to start by defining the latent trait related to the person and after that designing items to be

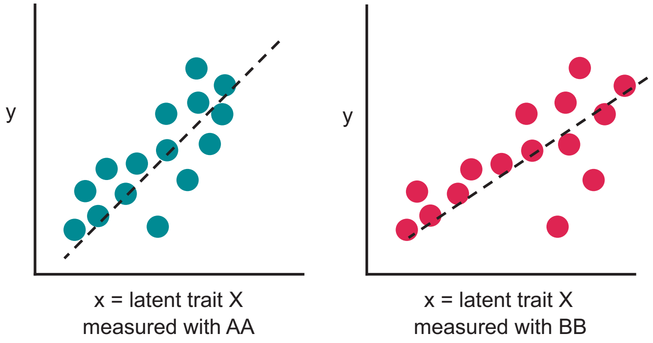


Figure 5.7: A fictive illustration for the same latent trait, *x*, measured with two different set of items that both fit the model well, correlated to *y*.

used to assess what it means for persons going up or down the scale (Wilson, 2005). However, historically, some psychologists have tended to view what is being measured as an empirical matter with a conception that views validity as something to be discovered afterward (Borsboom, Mellenbergh, & van Heerden, 2004). Furthermore, Fisher (1994) rhetorically asked *validity by default or design* and continued to claim that *just because experts have decided that items on a test all belong to the same content domain does not mean that they belong to the same construct*. Therefore, although having experts provide their views on what it means to go up or down the scale is a critical starting point, it is not enough to claim validity in either the latent trait itself or the latent trait as measured. Likewise, empirical data fitting the measurement model may help understand the latent trait itself and design measurements. Again, it is not enough in itself to claim validity in either the latent trait itself or the latent trait as measured.

As a key aspect when measuring latent traits, Morel and Cano (2017) stressed that *of all measurement properties, "content validity" is a sine qua non*. Indeed, rigorous research design, advanced statistics, nor large samples (Flake & Fried, 2020) can compensate for this afterward. A proper design to overcome limited – or, at worst misleading – measurements of a latent trait includes a *substantive patient-driven clinically anchored framework* (Morel & Cano, 2017), extending beyond health care in the human and social sciences. There is, however, an important differentiation between what is being used to measure a specific latent trait of interest (e.g., which items in a questionnaire) and what latent traits are of interest when making decisions. For the first point, setting up a set of items to be used to measure a specific latent trait attributed to the persons, must be carried out as a noncompetitive activity combining different expertise and resources (Morel & Cano, 2017). While, for instance, patients are key partners when developing measurements in health care, metrologists with expertise in latent traits must also be viewed as key partners. For example, they have a unique expertise in what requirements are important for designing good measurements. Therefore, they should be able to facilitate the work to hypothesize the composition of a latent trait of interest and how item can be mapped hierarchically along a clinical continuum (Barbic, Cano, & Mathias, 2018) and evaluate how well empirical data fit the measurement model as one source of information when developing measurements of latent traits. Designing items need to ensure enough variation in the item contents while at the same time staying within one dimension, and by asking enough questions to reduce uncertainty in relation to that variation (Fisher, Melin, & Möller, 2021). Furthermore, a good example of the latter comes from Morel and colleagues (2022), who provided a conceptual model for experiences from the early stage of Parkinson's disease, and in turn, lay the foundation of what latent trait to be measured in order to make a decision based on what matters for that target group.

The literature about designing measurements for latent traits, again naturally, starts by defining the latent trait related to the person and, after that, designing items (Wilson, 2005). A subsequent step when intending to test and evaluate the measurement in research is the study design. Much of that is, however, also a part of the description of the full measurement system, equally important to be considered and optimized in all measurement situations to give as good as possible measurements of the coupled latent traits. Recalling Figure 5.4, which shows a complete picture of the measurement process (Pendrill, 2019; Pendrill, 2014; Bentley, 2005) for latent traits, measurement information is transmitted from the measurement *object,* via an *instrument*, to an *operator* in the observation phase, which both the *environment* and the *measurement method* can influence. Design aspects of the operator, environment, and method also become apparent. Table 5.1 presents the measurement system components, entities for latent traits in general, and an example with memory measurements (Melin & Pendrill, 2022a).

Koopmans (1947) contrasts Tycho Brahe and Johannes Kepler's work nicely, where Brahe took a systematic approach of careful measurements, while Kepler looked for new models and was able to find simple empirical "laws" which were in accord with past observations as well as permitting the prediction of future observations. Combining theory-driven designs with an open-minded, explorative approach in an iterative and cross-disciplinary environment may foster the curiosity needed when new units to extend the SI are being explored. For instance, items – or persons – that do not fit the model may indicate multidimensionality and candidates for modification or discarding a theory (Baghaei, 2008). Likewise, items that do not match the expected hierarchy from theory or previous studies may warrant theory refinements (Karlsson et al., 2023). In particular, such anomalies will guide when and where to look for new phenomena (Kuhn, 1977) and pieces in the understanding of both the latent trait itself and the latent trait as measured. In turn, this need for iterative work will present possibilities for the design of measurements in the human and social sciences (Fisher & Stenner, 2011).

Table 5.1: Entities in the measurement system in general, for latent traits in general and exemplified for memory measurements.

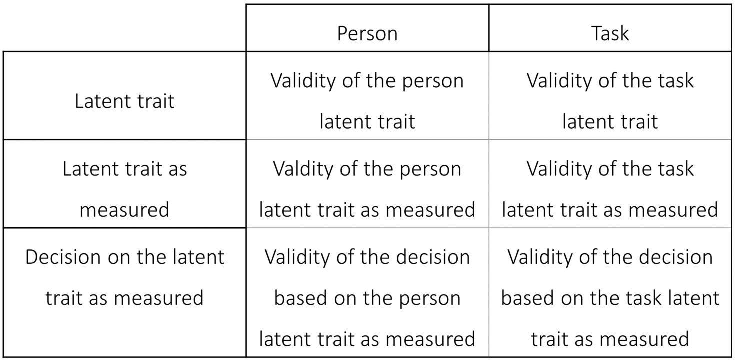
MSA term	Entities for latent traits in generall	Entities for example for memory measurements
Object	One test item	A sequence of numbers to be recalled
Objects	A set of test items	Sequences of numbers to be recalled
Instrument	One person	One person whose memory is being measured
Instruments	A cohort of persons	A cohort of persons whose memories are being measured
Operator, example role 1	The test leader	Study nurse
Operator, example role 2	The person observing	Study nurse
Operator, example role 3	The person doing the measurand restitution	Person doing the Rasch-analysis
Environment	The context of the measurement	Time on the day, place for testing
Method	Specifications in the observation phase	Item order, presentation of items
	Specifications in the measurand restitution phase	The measurand restitution

5.3.5 Construct specification equations as a mean of validity

Despite the fact that observed data can be provisionally and initially validated to a limited degree simply via fit to a probabilistic conjoint measurement model (eq. (5.1)), a measurement that lacks a construct theory is, as stated by Stenner et al. (2013), *a black box in which understanding may be more illusory than not*. Thus, more is needed to claim validity than just the fit of the data to a model and the demonstration of group invariance. In line with that, Stenner and colleagues (1982; 1983) introduced the so-called construct specification equations (CSEs) for latent traits attributed to tasks, which along with related approaches to devising explanatory measurement models (De Boeck & Wilson, 2004; Embretson, 2010; Fischer, 1973) are frequently stressed as a mean of validity when measuring latent traits of persons (McKenna et al., 2019; Stenner et al., 2006; Stenner et al., 2013; McKenna, Heaney, & Wilburn, 2019). Specifically, the CSE approach has to date been used mostly to explain the latent trait itself for tasks as an argument for the validity of the measured person attributes:

> *The rationale for giving more attention to variation in item scores is straightforward. Just as a person scoring higher than another person on a set of items is assumed to possess more of the construct in question (i.e., visual memory, reading comprehension, anxiety), an item (or task) that scores higher (in difficulty) than another item must be viewed as demanding more of the construct. The key question deals with the nature of the "something" that causes some persons to score higher than others and some items to score higher than others.* [. . .] *Such an equation embodies a theory of item-score variation. It simultaneously provides a straightforward means of confirming or falsifying alternative theories about the meaning of scores generated by a measurement procedure.* (Stenner & Smith 1982)

While Stenner and colleagues made a substantial contribution to advancing methods for validation – including CSE – of measurements in the human and social sciences, statements such as *an instrument is valid if it measures the intended attribute*, and *the "validity" concept should be equally applicable, to both attributes* (Stenner, 2014) is, however, somewhat contradicting. It may, to some extent, be explained by the lack of separating the latent trait itself and the latent trait as measured, as well as an improper description of the measurement system. Combining the fact that validity is applicable for both latent traits as well as for both the latent trait itself and the latent trait as measured is summarized in the matrix in Figure 5.8. Even though the main interest of the end-user is most often associated with making decisions about a latent trait attributed to the person, decisions attributed to the tasks are also applicable.[5]



	Person	Task
Latent trait	Validity of the person latent trait	Validity of the task latent trait
Latent trait as measured	Valdity of the person latent trait as measured	Validity of the task latent trait as measured
Decision on the latent trait as measured	Validity of the decision based on the person latent trait as measured	Validity of the decision based on the task latent trait as measured

Figure 5.8: The distinction of validity claims for the latent trait, the latent trait as measured, and the decision based on the latent trait as measured separated for latent traits attributed to persons and tasks.

We would argue for the CSE itself to be a "model" of the latent trait itself and consequently as a means of validity. In turn, the validity of the measured latent trait can be obtained when there is a high correlation between the measured latent trait and the latent trait itself, which applies for both coupling attributes such as person ability, θ, and

[5] An example of where a decision about the task(s) is the main interest is psychophysics, where a test panel is used to quantify specific latent traits attributed to products.

task difficulty, δ. In line with Stenner and colleague's work, we have also initially focused on CSEs for task difficulty (Melin et al., 2019, 2022a, 2022b; Pendrill, 2019; Melin, Cano, & Pendrill, 2021; Melin & Pendrill, 2023), although not only as a means for the validity of the measured latent trait of the person. Rather, the point of departure for CSEs for task difficulty has mainly been driven by the measurement system approach (presented in Section 2.2) where the human responder acts as the instrument (Pendrill, 2014). Specifically, we have suggested that CSEs *appear to provide metrological references for calibration and subsequent inter-comparability of measurements* (Melin et al., 2022b). Particularly, CSE can not only serve as a means of validity but also resemble formulas for "reference measurement procedures" (RMPs) analogous to RMPs found in the metrology of chemistry.

While the pure theoretical definitions of a latent trait attributed to a person θ_i can be defined as $g(\chi_i, \varepsilon_i, \beta)$ and a latent trait attributed to a task δ_j can be defined as $f\left(\upsilon_j, \epsilon_j, \gamma\right)$ (Figure 5.2), the CSE, however, can be considered a quasi-theoretical model of the latent trait itself:

$$\hat{\mathbf{Z}} = \sum_k \beta_k \cdot x_k \tag{5.2}$$

where $\boldsymbol{Z}$ is the latent trait of interest. In turn, $\mathbf{Z}$ is defined as a linear combination of a set, $\boldsymbol{k}$, (independent) variables, $\boldsymbol{X}$. Similar to a purely theoretical model, the CSE is equally applicable to both latent traits (Figure 5.9). Thus, some variables that cause variation in the latent trait attributed to persons explain why some people have better abilities than others. Likewise, some variables that cause variation in the latent trait attributed to tasks explain why some tasks are easier than others.

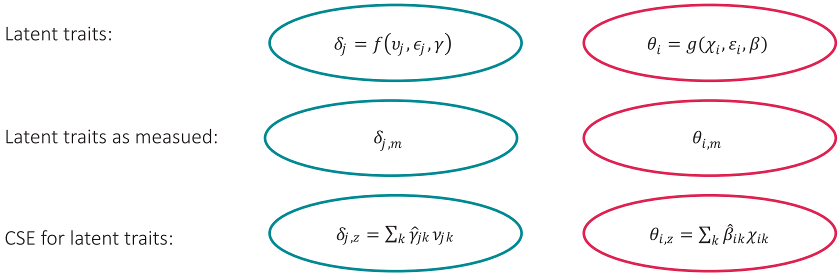

Latent traits:
$\delta_j = f\left(\upsilon_j, \epsilon_j, \gamma\right)$
$\theta_i = g(\chi_i, \varepsilon_i, \beta)$
Latent traits as measued:
$\delta_{j,m}$
$\theta_{i,m}$
CSE for latent traits:
$\delta_{j,z} = \sum_k \hat{\gamma}_{jk} \nu_{jk}$
$\theta_{i,z} = \sum_k \hat{\beta}_{ik} \chi_{ik}$

Figure 5.9: Notations separated for latent traits themselves, latent traits as measured, and CSE for latent trait separated for latent traits attributed to tasks δ_j and persons $\theta_{i.}$

In the EMPIR projects, NeuroMET 15HLT04 and NeuroMET2 18HLT09, researchers from national metrology institutes, academia, and industry have worked together to improve measurements for neurodegenerative diseases (Quaglia et al., 2021). One work package has been dedicated to memory measurements, which is one of the first metrological projects in European level to include measurements of latent traits. In the development of the NeuroMET Memory Metric, CSEs have been used as means of validity claims when combining different items from legacy tests. Block and digit recalling items reveal almost identical CSEs (Melin, Cano, & Pendrill, 2021), and two kinds of word recalling items reveal almost identical CSEs (Melin et al., 2022a; Melin & Pendrill, 2022). Furthermore, with entropy – originated from the Brillouin expression – dominating all CSEs (Melin et al., 2022b), they add validity that goes beyond a good fit to a measurement model (Melin et al., 2023a). Even though corresponding CSEs – including the dominating entropy contribution – for backward recalling block and number tasks have also been studied (Melin et al., 2023b), in the NeuroMET Memory Metric, only forward recalling sequences are included. This is because indications were found of multidimensionality when combining forward and backward recalling sequences as well as the set of items challenging the test person in terms of maintenance or manipulation working memory, respectively, and the constructs appeared less related and more likely to represent different underpinning constructs (Melin et al., 2023b). Thus, as argued above, fit statistics and a qualitative understanding are important, and this needs to go hand in hand also with the CSE.

An important note is that our CSE approach differs from the earlier work by Stenner and colleagues (1982; 1983) in choosing a principal component regression (PCR) rather than a regression based on the explanatory variables. This is because we cannot be sure *how* independent the explanatory variables are and whether they are the experimentally observed quantities or not (Melin & Pendrill, 2023). Specifically, when applying a principal component analysis in the PCR, one identifies the main components of variation by "rotating" in the explanatory variable space from the experimental dimensions to the principal component dimensions. Thus, when using principal components, we can allow some combination of the explanatory variables in cases where there is a significant correlation between them. A second important note is why we consider the CSE to be quasi-theoretical. This is because the linear regression is being made of the latent traits as measured – for persons $\theta_{i,m}$ or tasks $\delta_{j,m}$ – against $\boldsymbol{X}'$ in terms of the principal components. Ideally, we would have made the regression of the latent trait themselves, but obviously, it is not accessible. Thus, the measurement values of the latent trait are the closest to being used. It should, however, be noted that it warrants a good qualitative understanding of what is being measured and a good fit to the model to avoid developing misleading CSEs. On the other hand, a CSE may not only serve as a means of achieving validity, but as will be discussed in Section 5.4, it may also be used as an explorative tool when advancing the understanding of both the latent trait itself and the latent trait as measured when positioning models, measurement, and metrology to extend the SI.

5.3.6 Is validity straightforward or complex?

To close this section, we pick up on Borsboom et al. (2004), who claimed that *validity is not complex, faceted, or dependent on nomological networks and the social consequences of testing.* We agree that the concept's meaning can be very straightforward; nevertheless, the use of it has not been straightforward. Consequently, while validity has multiple meanings in measurement in the human and social sciences, a first step toward a more unified view of validity must separate the three validity claims presented in Figure 5.5, where the validity of the latent trait itself is often (or perhaps always?) a precondition to the validity of the latent trait as measured, which is often (or perhaps always?) in turn a precondition for the validity of the decision based on the latent trait as measures.

5.4 Routes to a better use of validity terminology and processes when extending the SI

Despite the fragmented use of validation processes, all agree that validity should be optimized. This work deals with it indirectly, but our key message calls for a better – optimized and clearer – terminology for validity. In turn, we believe that it will advance the validity of the latent traits themselves, the validity when measuring latent traits, and the validity in decisions about latent traits. Thus, the most important message is to understand the difference between the latent trait itself (Section 5.2.1) and the latent trait as measured (Section 5,2.2), and consequently, three important claims of validity need to be distinguished (Figure 5.5).

Furthermore, in a time when both latent traits ought to be understood, ways to measure the latent traits, and finding methods for validation are needed, we would encourage an iterative and cross-disciplinary approach rather than a too strict process. This is expected to advance the field of measurements in the human and social sciences when extending models, measurement, and metrology of the SI. At the same time, one must be careful not to make too strong validity claims.

5.4.1 Iterative, explorative, and cross-disciplinary efforts when measuring latent traits

On the one hand, clearly and consensus-based “rules” for validity in the human and social sciences when extending models, measurement, and metrology of the SI will support a more “production-like” process. On the other hand, it might be that the field is not yet ready for a “one-size-fits-all” approach. Of course, again, there is a need to have a harmonized view of terminology and possible limitations in claims at different stages. For instance, a too-hardline data-driven approach could be dangerous (Morel & Cano, 2017). Even when items are designed with a construct theory in mind, it might happen that observations do not vary as expected or do not fit the measurement model. For instance, uninterpretable inconsistencies might be due to an underdeveloped theory and/or low-quality data (Fisher, Melin, & Möller, 2021), but this should not be *a sign of the end of the conversation or of the measurement effort* (Fisher & Stenner, 2011).

Rewriting and/or changing items is the most common way of addressing the issue with the misfit. However, other aspects might also be related to the measurement system and the measurement process affecting model fit to be considered, adjusted for, and re-evaluated. For instance, how did the test leader interact with the test person, when and where was the observation, and what kind of specifications were used in the measurement restitution (Table 5.1)? As an example of the latter, specific objectivity (Rasch, 1966) is a unique feature of probabilistic conjoint models requiring separable parameters and minimally sufficient statistics, implying that the comparability of measurements of latent traits attributed to the person should be independent of which test items are being used and, symmetrically, comparability of item measures should be independent of which test persons are being used. In contrast with this capacity to support metrological traceability, other classes of models, such as those falling under the heading of "Item Response Theory" (IRT; Hambleton et al., 1991), cannot maintain unique metrological properties (Embretson, 1996; San Martin et al., 2009, 2015).

We do not assert the metrological viability of sociocognitive measurement without recognizing that

- local realizations and interpretations of even physical units of measurement may vary across communities of research and practice in divergent ways (Galison, 1997; Tal, 2014; Woolley & Fuchs, 2011);
- that irreducible randomness, incompleteness, and inconsistencies permeate elementary number theory, arithmetic, and Newtonian mechanics (Chaitin, 2003); and
- that longstanding calls for clearly distinguishing levels of complexity (Rousseau, 1985; Star & Ruhleder, 1996, p. 118) typically go unheeded.

We explicitly call for explorations of ways to separate levels of complexity in the measurement context and applaud recent efforts in this vein that expand on Galison's notion of the trading zone and Star's theory of the boundary object (Confrey et al., 2021; Fisher & Wilson, 2015; Lehrer & Jones, 2014). These efforts expand on Galison's (1997) documentation of the complex nonlinearities he found exhibited across the discontinuously interrelated microphysics communities of experimentalists, instrument makers, and theoreticians. Independent support for Galison's sense of the paradoxical positive functionality produced when convergent agreement is complemented by some kinds of divergent disagreement is provided by Woolley and Fuchs' (2011) study of collective intelligence in the organization of science.

Additional support is evident in Ostrom's theory of institutional organization, where a nested hierarchy of concrete operational rules, abstract collective-choice rules, and formal constitutional rules are distinguished: "Constitutional-choice rules affect operational activities and results through their effects in determining who is eligible and determining the specific rules to be used in crafting the set of collective-choice rules that in turn affect the set of operational rules" (Ostrom, 2015, p. 52; Kiser & Ostrom, 1982). We expect that our research results will make substantive contributions to furthering Ostrom's program of participatory social ecologies, in the manner described by Fisher and Stenner (2018).

We are especially focused on applications where significant portions of the population exhibit different sensitivities in discriminating differences (Melin et al., 2022a). Feedback on these differences may comprise concretely actionable information useful to end-users and so ought to be systematically reported to them in common languages and formats throughout interconnected, quality assured metrological networks (Fisher, Oon, & Benson, 2021; Mallinson, 2024; Penuel et al., 2016, 2020).

This methodology differs from that employed in IRT in that measurements are not assumed to reduce population characteristics in a homogenizing, deductive way, necessitating either the forcing of round pegs into square holes, or uncontrollable variation in unit definitions. Instead, because the measurement model is not meant to be true, but must be useful (Rasch, 1960, pp. 37–38; Rasch, 1972/2011; Box, 1979), and in accord with the idea that measurement extends and feeds back into everyday language (Fisher, 2020, 2023), standards are seen as mediating inherently unrealistic formal axioms and locally idiosyncratic concrete circumstances. We aim to revitalize dialogue at the point of use as a means by which ambiguities are reconciled and shared points of reference are negotiated, as when a request to "open a window" has to be clarified by mentioning the stuffy room, or pointing at a computer screen.

Mediating standards operationalize the substantive value and enhanced defensibility obtained vis-a-vis individualized inferences when theoretical explanations and empirical estimates of person and item locations are predictable, repeatable, and reproducible. The "black box" of empirical analyses demonstrating separable parameters in single instances lacking defined constructs is insufficient to the task of scientific measurement. Substantive understanding must be demonstrated via theoretical explanations and predictions.

The integration of formal and concrete levels of complexity in abstract measurements is then further augmented by restricting inferences so that the information represented is associated with and derived from the organizational level the data describe. We agree here with the hypothesis offered by Hayman, Rayder, Stenner, and Madey (1978, p. 31) that, "the closer a set of data is to the organizational level for which it will be used (for decision-making), the more useful the data will be." Thus, treating counts of correct responses or summed ratings as measurements commits the ecological fallacy (Alker, 1969; Gnaldi et al., 2018) by mistaking numbers for quantities (Fisher, 2021). Reporting only interval measurements to end-users invested in the concrete application of the original data then also obscures the very information on responses most vital to their decision processes.

Information on variation in item discrimination is not ignored at the abstract level of the measurements, of course, since it correlates very highly with commonly employed model fit statistics (see figure 2 in Wright, 1995) and can be reported for every item and every category transition threshold using software like Winsteps (Linacre, 2023). For examples of end-user reports illustrating statistical and graphical representations of individual anomalies, see figure 8.8.2 and table 8.8.1 in Wright and Stone (1979, pp. 207–208). Reporting concrete exceptions that prove the rule to end-users could productively complement the reporting of abstract SI unit measurements and formal explanatory CSE predictions.

In line with others, we have argued for the significance of proper designs of measurement. Although today's society already has huge access to data, it is expected to continue to increase. This opens up the need for even more explorative approaches to understanding new phenomena and networks. At the same time, even when taking a more explorative approach, neither latent traits themselves nor measurements of them happen purely by accident. It cannot only be an empirical matter with a conception that views validity as something to be discovered afterward. However, we can learn from empirical studies how these new ideas can be expressed in relation to existing ones (Fisher, Melin, & Möller, 2021). To quote Andrich and Marais (2019): *when the data do not accord with the model, then the model can still be very useful in understanding the data. It helps to diagnose where the data are different from what was expected from the model. Usually, there is an explanation for such effects.* Therefore, combining theory-driven designs with an open-minded explorative approach could help when seeking new units to extend the SI. For example, available data may be submitted to a measurement model-based analysis as a way of leveraging low-hanging fruit capable of indicating the possibility of defining a potential new item hierarchy, one that might consequently be rearticulated as a CSE.

CSEs might also be used as an explorative tool for advancing the understanding of a latent trait itself and, consequently, the latent trait as measured. A CSE provides a more specific, causal, and rigorously mathematical conceptualization of latent traits than any other construct theory (Melin, Cano, & Pendrill, 2021). Our previous work has suggested three key parts in selecting explanatory variables to be included in a CSE (Pendrill, 2019; Melin & Pendrill, 2023), which also can be seen as an exploration toward a better understanding of the latent trait itself. First, it must be a conceptual, practical, or clinical judgment to define appropriate explanatory variables to be tested. Secondly, statistical guidance is needed to find the most significant explanatory variables to be included in the CSE. For guidance, a univariate correlation study between the latent trait of interest and each explanatory variable is complemented, in the PCR, by a multivariate correlation matrix formulated to evaluate the degree of correlation and the intercorrelations between the explanatory variables. Thirdly, when developed in a PCR, the performance of the CSE itself, as well as the amount of contribution from each explanatory variable, to the latent traits as measured against $\boldsymbol{X}'$ in terms of the principal components is evaluated by (i) the strength of correlation between the prediction (i.e., $\theta_{i,z}$ or $\delta_{j,z}$) and the latent traits as measured (i.e., $\theta_{i,m}$ or $\delta_{j,m}$) and (ii) the dispersion of the β-coefficients of the CSE (Melin & Pendrill, 2023). Thus, by adding or removing variables, one can use the CSE as an exploratory tool to advance the understanding of the latent trait operationalized in the construct theory. Again, a good qualitative understanding of the latent trait to be measured and a good fit to the model are preconditions to avoid misleading CSEs and interpretations of what is causing variation in the latent trait itself attributed to a person $\theta_i, g(\chi_i, \varepsilon_i, \beta)$ and a latent trait attributed to a task $\delta_j, f(\upsilon_j, \epsilon_j, \gamma)$ (Figures 5.2 and 5.8).

Building on the work by Adroher and Tennant (2019), who used clinical judgments to explain variation in task difficulty in activities of daily living, and Fisher (2012), who rated variations in physical functioning items, CSEs formulated with *qualitative* explanatory variables may also be possible. In two recent studies, we have explored this for balance measurements (Melin et al., 2023c) and upper limb measurements (Wangdell et al., 2023). In both works, explanatory variables that linearly increase or decrease along the continuum of either balance task difficulty or upper limb task difficulty were identified. Importantly, one must seek the demands required for the tasks themselves, not for a specific person/group of persons performing them. In a second stage, healthcare professionals were invited to score each of the identified explanatory variables for each item in the Berg Balance Scale or Tetraplegia Upper Limb Questionnaire. Note that these were not the same healthcare professionals for both cases; they were recruited for each study with specific domain expertise, in a manner related to that described by Bunderson et al. (2009). Subsequent analysis of the scored explanatory variables provided estimates of linear interval measures for each variable that subsequently could be used in the CSE. While both studies have shown methodological and conceptual possibilities, several concerns to be considered in further work have been highlighted.

In both studies, we have also discussed the role of entropy as in our earlier NeuroMET studies. At the same time, one must remember that there must be general demands on the body to perform different tasks, which differs from explaining an individual person's ability (Melin et al., 2023; Wangdell et al., 2023). Secondly, one may use a group of people to define the explanatory variables; it is likely that a group of people who can all perform all tasks equally well will have a very low variation in an entropy measure, and the average entropy is expected to increase linearly with the difficulty of the tasks. We hope those works open for further discussion and investigations to advance measurement quality assurance by including CSEs as a means of validity for understanding the latent trait intended to be measured.

Finally, cross-disciplinary efforts when measuring latent traits are warranted. Potential key roles have been presented in Section 5.3.3, but we highlight the significance of developing structures and forums for such cross-disciplinary efforts here. For example, the EU-funded *Measuring the impossible* (Pendrill et al., 2010) could be seen as a forerunner where different disciplines met somewhere between psychology and engineering to advance measurement methods for the human and social sciences. Thus, a better understanding of validity when extending models, measurement, and metrology of the SI cannot be an isolated activity only within or only outside the metrological community.

5.4.2 Validity claims today and tomorrow

With exponential growth in society's need to make well-informed decisions based on latent traits, and at the same time, from a strict metrological perspective with undeveloped models and methods, the understanding of the latent trait themselves and practical tools and advanced methodologies to measure the latent traits must be developed simultaneously. Today, weaker validity claims than tomorrow must be allowed, and for some latent traits of interest, weaker validity claims than others must be allowed. Nevertheless, with weaker validity claims, this needs to be communicated transparently, and, in turn, responsibility must be taken for the consequences of decisions being made based on those claims.

Inspired by physical metrology and centuries of continuously improving the measurements, we must be dedicated to advancing methods for meeting society's needs for fit-for-purpose and high-quality measurements of latent traits. This includes not stopping with "our job" when finding a set of items fitting an appropriate measurement model. Rather, one must continue to test in new and diverse samples and cross-country studies, evaluate possibilities when adding items to improve targeting and reliability without jeopardizing validity, and so on. On a global level, this also includes establishing and coordinating metrological references to support comparable measurement values of latent traits over time or between different areas. Thus, by continuously improving our methods, tomorrow's validity claims will be stronger for latent traits.

5.5 Conclusion

When positioning models, measurement, and metrology to extend the SI, the concept of validity is essential. It is hoped that this review and discussion about validity and its related aspects in the human and social sciences will contribute to including a more nuanced view of validity. However, we have not provided and have no intention of providing a panacea or a one-size-fits-all route for better use of validity terminology and validity. Rather we have proposed different routes, originating from the three important claims of validity to distinguish, and we hope it will stimulate a fresh look at what might be possible.

Overall, claims about the validity of the latent traits themselves, the validity when measuring latent traits, and the validity in decisions about latent traits and methods for validation should not be mixed. Careful and responsible actions must be taken when making claims about measurement-related validity and decision-related validity before validity in the latent trait itself is ensured. However, it is important to remain open for an iterative, explorative, and cross-disciplinary effort to advance both the validity of the latent traits themselves and the measured latent traits and develop methods for validation of latent traits themselves and the measured latent traits.

References

Adroher, N. D., & Tennant, A. (2019). Supporting construct validity of the evaluation of daily activity questionnaire using linear logistic test models. *Quality of Life Research*, *28*(6), 1627–1639, https://doi.org/10.1007/s11136-019-02146-4

Alker, H. R. (1969). A typology of ecological fallacies. In M. Dogan & S. Rokkan (Eds.). *Quantitative ecological analysis in the social sciences* (pp. 69–86). MIT Press.

Andersen, E. B. (1999). Sufficient statistics in educational measurement. In G. N. Masters & J. P. Keeves (Eds.). *Advances in measurement in educational research and assessment* (pp. 122–125). Pergamon.

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*, 449–460.

Andrich, D. (1988). *sage university paper series on quantitative applications in the social sciences.* Vol. series no. 07–068: Rasch models for measurement. Sage Publications.

Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika*, *75*(2), 292–308.

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer Texts in Education, Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-7496-8

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions*, *22*(1), 1145–1146.

Barbic, S. P., Cano, S. J., & Mathias, S. (2018). The problem of patient-centred outcome measurement in psychiatry: Why metrology hasn't mattered and why it should. *Journal of Physics: Conference Series*, *1044*, 012069. https://doi.org/10.1088/1742-6596/1044/1/012069

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of pair comparisons. *Biometrika*, 63, 324–345.

De Battisti, F., Nicolini, G., & Salini, S. (2010). The Rasch model in customer satisfaction survey data. *Quality Technology & Quantitative Management. Taylor & Francis*, *7*(1), 15–34, https://doi.org/10.1080/16843703.2010.11673216

Bentley, J. P. (2005). *Principles of measurement systems*. 4th ed., Harlow, England ; New York: Pearson Prentice Hall.

Bohm, D. (1952). A suggested interpretation of the quantum theory in terms of "hidden" variables. I. *Physical Review*, *85*(2), 166–179.

Bohm, D. (2005). *Wholeness and the implicate order*. Routledge.

Bohm, D., & Hiley, B. J. (1984). Measurement understood through the quantum potential approach. *Foundations of Physics*, *14*(3), 255–274.

Bohm, D., & Hiley, B. J. (1989). Non-locality and locality in the stochastic interpretation of quantum mechanics. *Physics Reports*, *172*(3), 93–122.

Bohm, D., & Hiley, B. J. (2006). *The undivided universe: An ontological interpretation of quantum theory*. Routledge.

Bohm, D., Hiley, B. J., & Kaloyerou, P. N. (1987). An ontological basis for the quantum theory. *Physics Reports*, *144*(6), 321–375.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511490026

Denny, B., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219, https://doi.org/10.1037/0033-295X.110.2.203

Denny, B., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071, https://doi.org/10.1037/0033-295X.111.4.1061

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.). *Robustness in statistics* (pp. 201–235). Academic Press, Inc.

Bunderson, C. V., & Newby, V. A. (2009). The relationships among design experiments, invariant measurement scales, and domain theories. *Journal of Applied Measurement*, *10*(2), 117–137.

Chaitin, G. J. (1994). Randomness and complexity in pure mathematics. *International Journal of Bifurcation and Chaos*, *4*(1), 3–15, http://www.worldscientific.com/doi/pdf/10.1142/S0218127494000022

Chaitin, G. J. (2003). The limits of mathematics. Springer-Verlag.

Clifton, J. D. W. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods*, *25*(3), 259–270, https://doi.org/10.1037/met0000236

Confrey, J., Shah, M., & Toutkoushian, E. (2021). Validation of a learning trajectory-based diagnostic mathematics assessment system as a trading zone. *Frontiers in Education: Assessment, Testing and Applied Measurement*, *6*(654353), doi:10.3389/feduc.2021.654353

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin. US: American Psychological Association*, *52*, 281–302. https://doi.org/10.1037/h0040957

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Statistics for Social and Behavioral Sciences. Springer-Verlag.

Dybkaer, R. (2010). ISO terminological analysis of the VIM3 concepts 'quantity' and 'kind-of-quantity. *Metrologia*, *47*(3), 127, https://doi.org/10.1088/0026-1394/47/3/003

Embretson, S. E. (1996). Item Response Theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*(3), 201–212.

Embretson, S. E. (2010). *Measuring psychological constructs: Advances in model-based approaches*. American Psychological Association.

Engelhard, G., Jr (2012). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge Academic.

Finkelstein, L. (1975). Fundamental concepts of measurement: Definition and scales. *Measurement and Control*, *8*(3), 105–111. SAGE Publications Ltd, https://doi.org/10.1177/002029407500800305

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement*, *34*(1), 39–48, https://doi.org/10.1016/S0263-2241(03)00018-6

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Fisher, W. P., Jr (1988). *Truth, method, and measurement: The hermeneutic of instrumentation and the Rasch model*. Dissertation, Dissertation Abstracts International (University of Chicago, Dept. of Education, Division of the Social Sciences). 49, 0778A.

Fisher, W. P., Jr. (1994). The Rasch debate: Validity and revolution in educational measurement. In, 36–72.

Fisher, W. P., Jr (2020). Contextualizing sustainable development metric standards: Imagining new entrepreneurial possibilities. *Sustainability*, *12*(9661), 1–22, https://doi.org/10.3390/su12229661

Fisher, W. P., Jr (2021). Bateson and Wright on number and quantity: How to not separate thinking from its relational context. *Symmetry*, *13*(1415), https://doi.org/10.3390/sym13081415

Fisher, W. P., Jr (2023). Measurement systems, brilliant results, and brilliant processes in healthcare: Untapped potentials of person-centered outcome metrology for cultivating trust. In W. P. Fisher Jr. & S. Cano (Eds.). *Person-centered outcome metrology* (pp. 357–396). Springer, https://link.springer.com/book/10.1007/978-3-031-07465-3

Fisher, W. P., Jr (2012). A predictive theory for the calibration of physical functioning patient survey items. *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.2084490

Fisher, W. P., Jr, Melin, J., & Möller, C. (2021). *Metrology for climate-neutral cities*. http://urn.kb.se/resolve?urn=urn:nbn:se:ri:diva-57281. (12 September, 2022).

Fisher, W. P., Jr., Oon, E. P.-T., & Benson, S. (2021). Rethinking the role of educational assessment in classroom communities: How can design thinking address the problems of coherence and complexity? *Educational Design Research*, *5*(1), 1–33.

Fisher, W. P., Jr, & Jackson Stenner, A. (2011). Integrating qualitative and quantitative research approaches via the phenomenological method. *International Journal of Multiple Research Approaches*, *5*(1), 89–103, https://doi.org/10.5172/mra.2011.5.1.89

Fisher, W. P., Jr, & Jackson Stenner, A. (2018). Ecologizing vs modernizing in measurement and metrology. *Journal of Physics Conference Series*, *1044*(012025), http://iopscience.iop.org/article/10.1088/1742-6596/1044/1/012025

Fisher, W. P., Jr, & Wilson, M. (2015). Building a productive trading zone in educational assessment research and practice. *Pensamiento Educativo: Revista de Investigacion Educacional Latinoamericana*, *52*(2), 55–78, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2688260

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. In *Advances in methods and practices in psychological science*. SAGE Publications Inc. https://doi.org/10.1177/2515245920952393

Galison, P. (1997). *Image and logic: A material culture of microphysics*. University of Chicago Press.

Geisinger, K. F. (1992). The metamorphosis to test validation. *Educational Psychologist*, *27*(2), 197–222, https://doi.org/10.1207/s15326985ep2702_5

Gnaldi, M., Tomaselli, V., & Forcina, A. (2018). Ecological fallacy and covariates: New insights based on multilevel modelling of individual data. *International Statistical Review*, *86*(1), 119–135.

Esfeld, M., Lazarovici, D., Hubert, M., & Dürr, D. (2014). The ontology of Bohmian mechanics. *The British Journal for the Philosophy of Science*, *65*(4), 773–796, http://www.jstor.org/stable/24562842

Goldstein, S. (1998a). Quantum theory without observers-Part one. *Physics Today*, *51*(3), 42–47.

Goldstein, S. (1998b). Quantum theory without observers-Part two. *Physics Today*, *51*(4), 38–42, https://doi.org/10.1063/1.882241

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing: *Measurement and Evaluation in Counseling and Development*. *36*(3), 181–191. *Routledge*, https://doi.org/10.1080/07481756.2003.11909741

Hambleton, R. K., Swaminathan, H., & Rogers, L. (1991). *Fundamentals of item response theory*. Sage Publications.

Harvey, P., Jensen, C. B., & Morita, A. (2017). *Infrastructures and social complexity: A companion*. Taylor & Francis.

Hayman, J., Rayder, N., Stenner, A. J., & Madey, D. L. (1979). On aggregation, generalization, and utility in educational evaluation. *Educational Evaluation and Policy Analysis*, *1*(4), 31–39.

JCGM [20]0. (2012). *International vocabulary of metrology – Basic and general concepts and associated terms (VIM)*. BIPM.

Johansson, M., Preuter, M., Karlsson, S., Möllerberg, M.-L., Svensson, H., & Melin, J. (2023). *Valid and Reliable? Basic and Expanded Recommendations for Psychometric Reporting and Quality Assessment*. OSF Preprints, https://doi.org/10.31219/osf.io/3htzc

Joint Committee on the Standards for Educational and Psychological Testing. (2014). *Standards for Educational and Psychological Testing*. Washington, D.C: American Educational Research Association.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 198–211, https://doi.org/10.1080/0969594X.2015.1060192

Karlsson, S., Melin, J., Svensson, H., & Wisén, J. (2023). *A metrological approach to social sustainability metrics in municipalities*. OSF Preprints, https://doi.org/10.31219/osf.io/sdzwn

Kelley, T. L. (1927). *Interpretation of educational measurements.* (Interpretation of Educational Measurements). Oxford, England: World Book Co.

Kiser, L. L., & Ostrom, E. (1982). The three worlds of action: A metatheoretical synthesis of institutional approaches. In E. Ostrom (Ed.). *Strategies of political inquiry* (pp. 179–222). Sage.

Koopmans, T. C. (1947). Measurement without theory. *The Review of Economics and Statistics*, *29*(3), 161, https://doi.org/10.2307/1928627

Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change*. Revised. edition, Chicago, Ill: University of Chicago Press.

Lehrer, R. (2013). A learning progression emerges in a trading zone of professional community and identity. *WISDOMe Monographs*, *3*, 173–186.

Lehrer, R., & Jones, S. (2014, April 2). Construct maps as boundary objects in the trading zone. In W. P. Fisher Jr. (Chair). *Session 3-A: Rating scales and partial credit, theory and applied*. Philadelphia, PA: International Objective Measurement Workshop.

Linacre, J. M. (1995). Paired comparisons with ties: Bradley-Terry and Rasch. *Rasch Measurement Transactions*, *9*(2), 425, http://www.rasch.org/rmt/rmt92d.htm

Linacre, J. M. (2000a). Was the Rasch model almost the Peirce model? *Rasch Measurement Transactions*, *14*(3), 756–757, http://www.rasch.org/rmt/rmt143b.htm

Linacre, J. M. (2000b). Almost the Zermelo model? *Rasch Measurement Transactions*, *14*(2), 754, http://www.rasch.org/rmt/rmt142k.htm

Linacre, J. M. (2023). *A user's guide to WINSTEPS Rasch-Model computer program, v. 5.6.2*. Winsteps.com, https://www.winsteps.com/manuals.htm

Lovejoy, D. (1999). Objectivity, causality, and ideology in modern physics. *Science & Society*, *63*(4), 433–468.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review, 66*(2), 81–95.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new kind of fundamental measurement. *Journal of Mathematical Psychology*, 1(1), 1–27.

Mallinson, T. (2024). Extending the justice-oriented, anti-racist framework for validity testing to the application of measurement theory in re(developing) rehabilitation assessments. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*. in press, De Gruyter.

Mari, L., Wilson, M., & Maul, A. (2022). *Measurement across the sciences: Developing a shared concept system for measurement*. 1st ed. 2021 edition, Springer.

Matarese, V. (2023). *Epistemic studies. Vol. 51: The metaphysics of Bohmian mechanics: A comprehensive guide to the different interpretations of Bohmian ontology*. M. Esfeld, S. Hartmann, & A. Newen (Eds.). De Gruyter.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51–69, https://doi.org/10.1080/15366367.2017.1348108

Maul, A., Torres Irribarra, D., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement*, *79*, 311–320. https://doi.org/10.1016/j.measurement.2015.11.001

McAllister, S. (2008). Introduction to the use of Rasch analysis to assess patient performance. *International Journal of Therapy & Rehabilitation*, *15*(11), 482–490. Mark Allen Holdings Limited, https://doi.org/10.12968/ijtr.2008.15.11.31544

McKenna, S. P., Heaney, A., & Wilburn, J. (2019). Measurement of patient-reported outcomes. 2: Are current measures failing us? *Journal of Medical Economics*, *22*(6), 523–530, https://doi.org/10.1080/13696998.2018.1560304

McKenna, S. P., Heaney, A., Wilburn, J., & Jackson Stenner, A. (2019). Measurement of patient-reported outcomes. 1: The search for the Holy Grail. *Journal of Medical Economics*, *22*(6), 516–522. Taylor & Francis, https://doi.org/10.1080/13696998.2018.1560303

Melin, J. (2021). Neurogenerative disease metrology and innovation: The European Metrology Programme for Innovation & Research (EMPIR) and the NeuroMET projects. Conference presentation presented at the Pacific Rim Objective Measurement Symposium 2021. https://proms.promsociety.org/2021/.

Melin, J., Cano, S., Flöel, A., Göschel, L., & Pendrill, L. (2022a). The role of entropy in construct specification equations (CSE) to Improve the validity of memory tests: Extension to word lists. *Entropy*, *24*(7), 934. Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/e24070934

Melin, J., Cano, S. J., Flöel, A., Göschel, L., & Pendrill, L. R. (2022b). Metrological advancements in cognitive measurement: A worked example with the NeuroMET memory metric providing more reliability and efficiency. *Measurement: Sensors*, 100658. https://doi.org/10.1016/j.measen.2022.100658

Melin, J., Cano, S. J., Gillman, A., Marquis, S., Flöel, A., Göschel, L., & Pendrill, L. R. (2023a). Traceability and comparability through crosswalks with the NeuroMET memory metric. *Scientific Reports*, *13*(1), 1–12. Nature Publishing Group, https://doi.org/10.1038/s41598-023-32208-0

Melin, J., Pendrill, L. R., & Cano, S. J. EMPIR NeuroMET 15HLT04 consortium. (2019) Towards patient-centred cognition metrics. *Journal of Physics: Conference Series*, 012029. https://doi.org/10.1088/1742-6596/1379/1/012029

Melin, J., Cano, S., & Pendrill, L. (2021). The role of entropy in construct specification equations (CSE) to improve the validity of memory tests. *Entropy*, *23*(2), 212. Multidisciplinary Digital Publishing Institute https://doi.org/10.3390/e23020212

Melin, J., Göschel, L., Hagell, P., Westergren, A., Flöel, A., & Pendrill, L. (2023b). Forward and backward recalling sequences in spatial and verbal memory tasks: What do we measure? *Entropy*, *25*(5), 813. Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/e25050813

Melin, J., Fridberg, H., Ekvall Hansson, E., Smedberg, D., & Pendrill, L. (2023c). Exploring a new application of construct specification equations (CSEs) and entropy: A pilot study with balance measurements. *Entropy*,

Melin, J., & Pendrill, L. (2022a). Humans as measurement instruments and Construct specification equations (CSE) in measurement systems. *BEAR Seminar*, https://files.bearcenter.org/video/Melin Pendrill_HumansEquationsMeasurementSystems_20221115.mp4 ((4 May, 2023)).

Melin, J., & Pendrill, L. (2022b). A novel metrological approach to a more consistent way of defining and analyzing memory task difficulty in word learning list tests with repeated trials. In *Proceedings of the RaPID Workshop – Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments – Within the 13th Language Resources and Evaluation Conference* (pp. 17–21). Marseille, France: European Language Resources Association, https://aclanthology.org/2022.rapid-1.3 13 January, 2023.

Melin, J., & Pendrill, L. R. (2023). The role of construct specification equations and entropy in the measurement of memory. In F. William P. Jr. & S. J. Cano (Eds.). *Person-centered outcome metrology: principles and applications for high stakes decision making (*Springer series in measurement science and technology) (pp. 269–309). Cham: Springer International Publishing, https://doi.org/10.1007/978-3-031-07465-3_10

Messick, S. (1989a). Validity. In *Educational measurement*.3rd ed. (The American council on education/ Macmillan series on higher education) (pp. 13–103). American Council on Education.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. American Educational Research Association, https://doi.org/10.3102/0013189X018002005

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23. American Educational Research Association, https://doi.org/10.3102/0013189X023002013

Messick, S. (1996). Validity and Washback in Language Testing. *ETS Research Report Series*, *1996*(1), i–18, https://doi.org/10.1002/j.2333-8504.1996.tb01695.x

Michell, J. (2021). “The art of imposing measurement upon the mind”: Sir Francis Galton and the genesis of the psychometric paradigm. *Theory & Psychology*, *26*.

Morel, T., & Cano, S. J. (2017). Measuring what matters to rare disease patients – Reflections on the work by the IRDiRC taskforce on patient-centered outcome measures. *Orphanet Journal of Rare Diseases*, *12*(1), 171, https://doi.org/10.1186/s13023-017-0718-x

Morel, T., Cleanthous, S., Andrejack, J., Barker, R. A., Blavat, G., Brooks, W., Burns, P., et al. (2022). Patient experience in early-stage Parkinson’s Disease: Using a mixed methods analysis to identify which concepts are cardinal for clinical trial outcome assessment. *Neurology and Therapy*, *11*(3), 1319–1340. https://doi.org/10.1007/s40120-022-00375-3

Lim, M., Sok, S. R., & Brown, T. (2009). Using Rasch analysis to establish the construct validity of rehabilitation assessment tools. *International Journal of Therapy and Rehabilitation*, *16*(5), 251–260, https://doi.org/10.12968/ijtr.2009.16.5.42102

Newby, V. A., Conner, G. R., Grant, C. P., & Bunderson, C. V. (2009). The Rasch model and additive conjoint measurement. *Journal of Applied Measurement*, *10*(4), 348–354.

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational & psychological assessment*. Vol. 55, London: City Road. https://doi.org/10.4135/9781446288856.

Ostrom, E. (2015). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press. Original work published 1990.

Peirce, C. S. (1878). Illustration of the logic of science. Fourth paper: The probability of induction. *Popular Science Monthly*, *12*, 705–718. (Rpt N. Houser & C. Kloesel (Eds.). 1992, *The essential Peirce: Selected philosophical writings, vol. I*. 1867–1893, (pp. 155–169), Indiana University Press.).

Pendrill, L. (2014). Man as a measurement instrument. *NCSLI Measure*, *9*(4), 24–35, https://doi.org/10.1080/19315775.2014.11721702

Pendrill, L. (2019). *Quality assured measurement: Unification across social and physical sciences*. (Springer Series in Measurement Science and Technology). Springer International Publishing, https://doi.org/10.1007/978-3-030-28695-8

Pendrill, L. (2018). Assuring measurement quality in person-centred healthcare. *Measurement Science and Technology*, *29*(3), 034003, https://doi.org/10.1088/1361-6501/aa9cd2

Pendrill, L. (2021). Quantities and units in quality assured measurement. Presented at the PACIFIC RIM OBJECTIVE MEASUREMENT SYMPOSIUM 2021. https://proms.promsociety.org/2021/.

Pendrill, L. (2024). Quantities and units: Order amongst complexity. In W. P. Fisher, Jr. & L. R. Pendrill (Eds.), *Models, measurement, and metrology extending the SI*, (pp. 35–100). De Gruyter.

Pendrill, L. R., Emardson, R., Berglund, B., Gröning, M., Höglund, A., Cancedda, A., Quinti, G., et al. (2010). Measurement with persons: a European network. *NCSLI Measure*, *5*(2), 42–54. Taylor & Francis, https://doi.org/10.1080/19315775.2010.11721515

Penuel, W. R., Clark, T. L., & Bevan, B. (2016). Infrastructures to support equitable STEM learning across settings. *After School Matters*, *24*, 12–20.

Penuel, W. R., Riedy, R., Barber, M. S., Peurach, D. J., LeBouef, W. A., & Clark, T. (2020). Principles of collaborative education research with stakeholders: Toward requirements for a new research and development infrastructure. *Review of Educational Research*, *90*(5), 627–674.

Prigogine, I. (1971). Unity of physical laws and levels of description. In I. Prigogine & M. Grene (Eds.). *Interpretations of life and mind: Essays around the problem of reduction* (pp. 1–13). Humanities Press.

Prigogine, I. (1976). Order through fluctuation: Self-organization and social system. In E. Jantsch & C. Waddington (Eds.). *Consciousness and evolution: Human systems in transition* (pp. 93–130). Addison Wesley.

Prigogine, I. (1978). Time, structure and fluctuations [Nobel lecture]. *Science*, *201*, 777–785.

Prigogine, I., & Stengers, I. (2018). *Order out of chaos: Man's new dialogue with nature*. Verso.

Quaglia, M., Cano, S., Fillmer, A., Flöel, A., Giangrande, C., Göschel, L., Lehmann, S., Melin, J., & Teunissen, C. E. (2021). The NeuroMET project: Metrology and innovation for early diagnosis and accurate stratification of patients with neurodegenerative diseases. *Alzheimer's & Dementia*, *17*(S5), e053655. John Wiley & Sons, Ltd, https://doi.org/10.1002/alz.053655

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.

Rasch, G. (1966). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.). *Readings in mathematical social science* (pp. 89–108). Science Research Associates, https://www.rasch.org/memo19662.pdf

Rasch, G. (1973/2011). All statistical models are wrong! Comments on a paper presented by Per Martin-Löf, at the Conference on Foundational Questions in Statistical Inference, Aarhus, Denmark. *Rasch Measurement Transactions*, *24*(4), 1309. May 7–12, 1973, http://www.rasch.org/rmt/rmt244.pdf

Rousseau, D. M. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. *Research in Organizational Behavior*, *7*(1), 1–37.

San Martin, E., Gonzalez, J., & Tuerlinckx, F. (2009). Identified parameters, parameters of interest, and their relationships. *Measurement: Interdisciplinary Research and Perspectives*, *7*(2), 97–105.

San Martin, E., Gonzalez, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3 PL model. *Psychometrika*, *80*(2), 450–467.

San Martin, E., & Rolin, J. M. (2013). Identification of parametric Rasch-type models. *Journal of Statistical Planning and Inference*, *143*(1), 116–130.

Slaney, K. (2017). *Validating psychological constructs*. London: Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-38523-9

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, *7*(1), 111–134.

Stenner, A. J. (2014). Validity revisited. *Presented at the IOMW – Philadelphia April*, *1*, 2014.

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement*, *7*(3), 307–322.

Stenner, A. J., Fisher, W. P., Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology*, *4*, https://doi.org/10.3389/fpsyg.2013.00536

Stenner, A. J., & Smith, M. (1982). Testing construct theories. *Perceptual and Motor Skills*, *55*(2), 415–426, https://doi.org/10.2466/pms.1982.55.2.415

Stenner, A. J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, *20*(4), 305–316.

Tal, E. (2014). Making time: A study in the epistemology of measurement. *The British Journal for the Philosophy of Science*, *67*(1), 297–335, https://doi.org/10.1093/bjps/axu037

Tesio, L. (2003). Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *Journal of Rehabilitation Medicine*, 2003, 11.

Tesio, L., Caronni, A., Kumbhare, D., & Scarano, S. (2023). Interpreting results from Rasch analysis 1. The “most likely” measures coming from the model. *Disability and Rehabilitation*, 1–13. https://doi.org/10.1080/09638288.2023.2169771

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, XXXIII, 529–544. (Rpt L. L. Thurstone. 1959, *The measurement of values.* (pp. 215–233). University of Chicago Press, Midway Reprint Series).

Van Iddekinge, Chad, H., Lievens, F., & Sackett, P. R. (2023). Personnel selection: A review of ways to maximize validity, diversity, and the applicant experience. *Personnel Psychology* n/a(n/a). https://doi.org/10.1111/peps.12578.

Johanna, W., Pendrill, L., Dunn, J., Hill, B., & Melin, J. (2023). Construct specification equations to improve validity in upper limb measurements. *Frontiers in Rehabilitation Sciences*. https://assets.researchsquare.com/files/rs-4128671/v1_covered_4172cb71-d2ec-41c1-8f72-f5ccc01a2afa.pdf

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, N.J: Lawrence Erlbaum Associates.

Wolf, M. G., Ihm, E., Maul, A., & Taves, A. (2019). Survey item validation. In M. Stausberg & S. Engler (Eds.). *Handbook of Research Methods in the Study of Religion*. Vol. 10. 2nd ed., Routledge.

Woolley, A. W., & Fuchs, E. (2011). Collective intelligence in the organization of science. *Organization Science*, *22*(5), 1359–1367.

Wright, B. D. (1995). 3 PL IRT or Rasch? *Rasch Measurement Transactions*, *9*(1), 408, http://www.rasch.org/rmt/rmt91b.htm

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*(4), 33–45, 52, https://doi.org/10.1111/j.1745-3992.1997.tb00606.x

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.

Zermelo, E. (1929). The calculation of tournament results as a maximum-likelihood problem [German]. *Mathematische Zeitschrift*, *29*, 436–460.

William P. Fisher Jr.

6 Measurement logic, aesthetics, and ethics across the sciences: extending the SI units

Abstract: Several decades of theory and practice suggest that the natural and social sciences could productively share a common conceptual frame of reference for measurement. Cautions in this regard are evident in James Scott's recounting of the long history of only partially successful, and often catastrophically failed, efforts aimed at improving the human condition. Scott defines the terms of an integrity test to which all proposed social systems should be subject. But even if that test is passed, further questions arise as to whether evidence, logic, and methods alone are sufficient to the task. What possibilities are there for transcending the merely utilitarian and functional with higher aspirations to ideals of meaning, beauty, and justice? Might a compelling case be made for captivating imaginations with opportunities for creative expression and for inspiring passionate engagement with problems of human suffering, social discontent, and environmental degradation? Answers to these questions begin from the widely recognized status of observers across fields from physics and psychiatry to ecology and anthropology as all participating in bringing phenomena into language as sharable objects of human interests. This perspective is substantiated by expanding on the physicist John Wheeler's sense of the participatory role the observer plays in manifesting phenomena of interest. The chapter concludes with three brief proposals to revisit major initiatives compromised by unexamined assumptions about measurement, with the goal of suggesting how an extended SI could inform new institutional infrastructures better equipped to fulfill their intended purposes.

Keywords: social ecology, participant observer, meaning, beauty, aesthetics, psychology, semiotics, language, modeling, physics, complexity, standards, ethics

6.1 Introduction

6.1.1 Grounding the apparently audacious in common sense

The primary purpose of this book is to initiate a serious discussion of the possibility that the SI units might be extended to include a new class of complex constructs that have been successfully measured across a wide range of natural and social sciences for several decades. These constructs have exhibited persistent log-interval quantita-

William P. Fisher Jr., BEAR Center, Berkeley School of Education, University of California, Berkeley; Living Capital Metrics LLC

https://doi.org/10.1515/9783111036496-006

tive properties and metrological viability across multiple studies over considerable spans of time and space. Though these constructs are not currently measured using instruments subject to consistent quality assurances traceable to defined unit quantities, theory and evidence strongly suggests they could be.

Scales of this kind have been developed and in use for several decades, often in high stakes and global contexts. Though the public and many experts in psychological and social research and practice remain unaware of what has been achieved, a wealth of predictive theory and experimental evidence supporting the notion of an expanded SI is available to anyone interested in exploring it. Approaching these accomplishments in the context of the history of measurement from ancient times to the present, and in awareness of the immense learning facilitated by it, one has to be more than a little surprised to find that so little in the way of a serious conversation about extending the SI has occurred.

Of course, that surprise is tempered by awareness of the stark contrast between the complex, playful, and participatory worldview implied by advanced measurement modeling and the modern worldview's perspective on human subjectivity as alienated from the objectivity of an independently existing clockwork universe. Here, very good reasons can be discerned as to why important advances in more meaningful, useful, and communicable measurements could have been on the record for so long without being widely recognized for their value and implemented on broad scales.

The fact is that the kind of social and psychological measurement capable of informing an extended SI cannot be fit into the modern worldview. Obviously, the modern worldview already informs the existing paradigm of fragmented and incommensurable social and psychological measurement. Here, the assumption that science merely describes an independently existing world frames the idea that measurement assigns numbers to observations in ways that result in commensurable and additive interval units when those assignments involve objective reality. The ordinal incommensurability of social and psychological measurements, from this point of view, is simply a consequence of their subjective nature and nothing at all can be done to change this.

That modern worldview does not, however, provide a true account of measurement as it is practiced in either the natural or the social sciences. Decades of debate and reflection on the philosophy and history of science, measurement, and mathematics complement the results of science itself in pointing toward a different, unmodern or nonmodern worldview.

A nonmodern worldview recognizes humanity as participating in evolving and unfolding dialogues with nature and with itself. In the latter half of the twentieth century, physicists like David Bohm (1980), Ilya Prigogine (1976), and John Wheeler (1974) described a new kind of participant observer engaged in objectively repeatable, reproducible, playful, flowing, and dance-like relationships and interactions. Prigogine (1986, p. 494), for instance, explicitly seeks to recognize that we have been "victims of a distorted representation of science," and that a "new dialogue with nature" (Prigogine & Stengers, 1984) is underway.

Researchers in the philosophy, history, and social studies of science have similarly also articulated these themes, developing rich resources informing possibilities for transforming the conception and operationalization of mathematical models and methods. Toulmin (1982) and Gadamer (1976, 1989), for instance, both develop extensive arguments presenting method as fundamentally participatory, while Latour (1987, 2005) argues that passionate engagement in science is far less of a problem than ignoring the central importance of the propagation of representations across media. As he (Latour, 1996, p. 5) writes, "to believe that involvement, transformation, adulteration, reformatting and displacement weaken a 'Pure Science' of 'Pure Objectivity' is to have never seen a practicing scientist at work."

And so we see that a complex array of daunting conceptual and operational challenges is posed by the idea that the SI could be extended into new domains. Most of these issues are not addressed in the chapters of this book, which focus on presenting examples of the kinds of theories and evidence that support the feasibility, desirability, and viability of an extended SI unit system. But contrary to the first impulse likely felt by many upon encountering the idea of scientifically rigorous quantification of human and social values, many readers may find that the chapters in this book point toward clear possibilities for individually and globally meaningful systems of comparison, and that these are to be preferred over those that are equally meaningless at all scales.

So, the extended SI proposed in the title of this book entails neither an acceptance of ordinal level scores as measurements, nor a mere elevation of psychology and the social sciences to an epistemological equivalence with the natural sciences. Given that the very idea of an extended SI will strike many as highly implausible and beyond audacious as a serious proposal, new measurement capacities obtaining on a scale justifying an extended SI well ought to open up onto new, broader horizons and a transformed vision of human possibilities. This chapter is an initial attempt at locating these horizons and envisioning those possibilities.

6.1.2 Raising the level at which the intelligence of all operates

Decades of peer-reviewed research offer logical arguments and repeatedly reproduced scientific evidence showing that measurement theory and practice can be conceived in common terms across the natural and social sciences. Though necessary to the justification of an extended SI, these careful and detailed explanations and demonstrations have not led to a general consensus accepting and implementing a common frame of reference for measurement. Rational consideration of the weight of evidence, the effectiveness of theory, and the practical advantages to be gained would seem to lead inexorably to the expectation that intensive investments of resources in a new, extended array of SI units should have been imminently expected at any time over the last 50 years and more. But the complex, multilevel, and nonreductive logic involved, its mathematical derivations and forms, and the eminently practical and meaningful value of its results have not proven as persuasive as might be expected.

Why? Why has an extended SI failed to emerge? Why is it not even on the table as a point of discussion and debate? Possible answers to these questions take shape with the realization that cultures evolve via collective processes not directed by the conscious intentions of any individual or group, and that these processes moreover cannot be explained solely in rational, logical terms. This realization is in fact consistent with the complex issues involved in measurement since the premises of science and mathematics have long been recognized as relying on ineradicable presuppositions that are neither empirical nor logical (Burtt, 1954; Butterfield, 1957; Chaitin, 1994; Floyd & Kanamori, 2016; Gadamer, 1981; Holton, 1988, p. 41; Kuhn, 1970; Nagel & Newman, 1958, p. 101; Polanyi, 1974; Toulmin, 1953; Weinsheimer, 1985, p. 52; Wittgenstein, 1983).

To be convincing in compelling ways, then, measurement logic, tools, and methods may need to be complemented by other kinds of arguments and experiences. What relevant metaphors, for instance, might be capable of captivating imaginations and inspiring passions in new ways? How might aesthetics and ethics serve as complementary sources of rationales for improved measurement across the sciences?

The modern Cartesian sense of method as rules followed in the production of truth fails to account for the history of science (Kuhn, 1970, 1977; Richards & Daston, 2019; Toulmin, 1953). Scott (1998) documents how methodical efforts aimed at improving the human condition are typically blind to chaotic, emergent issues beyond the conceptual scope of modernist thinking. Scott (1998, pp. 355–357) then suggests everyday language as an alternative, as it offers a paradigmatically different sense of method, one that models ongoing collective processes of concept formation and virally communicable meaning in more fluid, humanly accessible ways. Here, phonemic, orthographic, and grammatical standards serve as the media through which shared understandings connect unrealistic formal ideals with unique local circumstances. Here, beauty, metaphor, ethics, and justice might more clearly motivate reasons for demanding new infrastructural standards for scientific, legal, market, and communications institutions.

So, what deeper sources of desire for beauty and meaning might be methodically accessible? How is that access acquired if it cannot be obtained via logic alone? What kind of a new system of root metaphors and world hypotheses (Pepper, 1942) might be productively imagined? We plainly must begin from understanding how the modern age and its postmodern deconstructions remain counterproductively enthralled with images of a clockwork universe, of humanity as the crown of creation, and of time as money. The popular press offers imaginative explorations of how humanity could experience itself as enchanted and at home in a living, evolving universe (Capra & Luisi, 2014; Cole, 1985; Kauffmann, 1996; Laszlo, 2019; Sahlins, 2022; Swimme & Tucker, 2011; Wheeler, 1994; Zukav, 1979), but these uniformly lack any specification of pragmatic methods for moving in new directions.

Most telling, perhaps, in this literature is the internal inconsistency displayed in the way arguments as to the truth and value of the idea that we live in a universe of interconnected wholeness are employed to educate and persuade individuals. Should not a science of that kind involve the methodical materialization of collective consciousness and its systematic deployment in an expanding network of traceable metrological connections? If we understand that undifferentiated wholeness is a real higher order of complexity fundamentally discontinuous with separate particularity, why would we expect individuals to be able to shift their consciousness into that sphere by a mere act of will? Won't it be necessary to create the cognitive infrastructure of a new SI unit system and associated vocabularies if we are to raise the intelligence of all in new domains without changing individual capacities "one whit" (Dewey, 1954, p. 210)? If we are to put reason on a higher imaginative level, do not better instruments offer more practical ways forward than trying to provide individuals with finer imaginations (Whitehead, 1925, p. 107)? Is not the point one of creating a "new social synthesis in which individuality and collectivity will not be exclusively opposed" (King, 1989, p. 46)? Should not we be "*thinking spirituality as infrastructural*" (Puig de la Bellacasa, 2015, p. 58; original emphasis) in the manner of Linssen's (1958, pp. 13–14) suggested Zen "spiritual materialism"? Is not the problem one of resolving the paradox of apparently opposed physical and social values (Overton, 2015, p. 43)? To reconcile humanity's global identity with each person's singular existence (Ricoeur, 1974a, p. 166)? Is not a metrology of extended SI units essential to developing models and methods "in which science and collective rationality may interact in a constructive way" (Prigogine, 1986, p. 504)?

Peirce (1955, p. 62; Peirce et al., 1977, pp. 85–86) can be read as suggesting that aesthetics and ethics must be semiotically integrated with logic, as pragmatic considerations focusing attention on how thinking and communication take place only in signs. That suggestion has to be situated in the context of one of the basic lessons taught by historical transformations of reason (Gadamer, 1979), which is that the rationality infusing logic and method does not primarily or originally emerge from within individual consciousness, intentionality, and rule-following. Instead, individual capacities are interwoven with the dominant paradigm's root metaphors (Black, 1962; Brown, 2003; Habermas, 1979; Hesse, 1970; Kuhn, 1993; Lakoff & Johnson, 1980, 1999; Ricoeur, 1977) and the available knowledge infrastructure's cognitive supports (Bateson, 1972; Hutchins, 2014; Petracca & Gallagher, 2020; Sutton et al., 2010; Vygotsky, 1978).

Pragmatic applications of this lesson ultimately focus on enhancements to knowledge infrastructures like the SI units. Seen in evolutionary terms, the focus of natural selection is on the unity of thoughtful individuals' cognitive and caring capacities with the opportunities for their actualization available in the external environment: flowers are to sunshine, earth, water, air, bees, and butterflies as thinking is to phonemes, alphabets, grammars, vocabularies, and unit standards (along with nutritious food, personal relationships, etc.) (Akrich & Latour, 1992; Bateson, 1972; Elsbach et al., 2005; R. Fisher, 1934a, p. 115; Heras-Escribano, 2020; Overton, 2002, 2007; Sutton et al., 2010; Watts, 1973). Taking the integrated organism-environment as the unit of natural selection in this way sets up a metasystematic level of hierarchical complexity for design thinking (Commons & Bresette, 2006; Commons et al., 2011). At this level, the focus of planning shifts from educating and persuading individuals and from centrally planned policies and programs to capacities for shaping cultural contexts serving as niches for new forms of social life (Akrich et al., 2002; Brown, 2008; Fisher & Stenner, 2018a; Hargadon, 2003; Latour, 1998; Law, 2009; Popescu, 2016; Wagner, 2023).

The idea here is to unite individual thinking with external standards by materializing collectively projected structural invariances in a common language of quality assured SI units. This theme is also taken up in the physics of nonequilibrium systems. Prigogine (1986, p. 504), for instance, says that his interest in this "class of models is that they enable us to make the interplay between the actors and the constraints of the environment more transparent." Prigogine overtly hopes models of this kind "will be a way of demythologizing the process of collective decision making, without negating its complexity."

Here, the role of language as a complex, multilevel model of the modeling process itself sets up ways of integrating logic with aesthetics and ethics, and of shifting the paradigm from the modern Western Cartesian assumption of subject-object dualism to an unmodern playful unity of subject and object absorbed together in the flow of experience. Though the challenges encountered here are daunting, paraphrasing Rasch (1960, p. xx), we can anticipate that, having formulated the problem, humanity will likely find a way to meet it.

6.2 Semiotics: taking language as a model for integrating individual and social minds

Natural language emerges as a model to follow in this effort, following on the insights offered by Scott (1998) at the conclusion of his close study of the history of "high modern" approaches to improving the human condition. Scott sees the problem in terms of language's capacity to integrate flexible local adaptations to concrete circumstances with abstract idealizations facilitating navigable continuity.

Two independent but parallel developments augment Scott's suggestion of natural language as a model. First, Star and Griesemer (1989), and Star and Ruhleder (1996), draw from Bateson's (1972) conceptualization of the problem of simultaneous concreteness and ideality as levels of learning obtaining across interrelated linguistic ecologies (Bowker et al., 2015). Second, Ostrom (2015, pp. 24–25), recognizing the need to create theoretical models more closely aligned with the empirical world, combines ideas from institutional economics and biologists' studies of nature to propose a theory of self-organizing and self-governing forms of collective action. Three-level ecological structures have also been adopted in public health as a way of productively separating and balancing efforts focused on within-individual micro, between-individual meso, and social macro-level processes (Bizouarn, 2016; Susser & Susser, 1996). These biosocial levels align well with the linguistic levels of complexity proposed by Bateson, which has led to research applying Star's concepts of ecologies of information infrastructure in Ostrom's sphere of participatory design principles for common pool resource management (Björgvinsson, 2014; Marttila, 2016; Karasti et al., 2016).

Attending to the quality of information and its communication across language's structural levels of complexity leads to an understanding of how living institutions can encompass both continuity and navigability, on the one hand, and openness and adaptability, on the other. Scientific support for democratic and economic freedom might then be extended into new domains by securing a basis for rights to the personal ownership of social and natural resources heretofore held only in common pools. The potential for mass customization of locally situated implementations of abstract ideals in the measurement and management of human, social, and natural capital sets up new horizons for leveraging the profit motive as the source of energy for driving a new economics of genuine wealth (Fisher, 2002, 2007, 2010a/c/d, 2011a, 2012b/c, 2020, 2023a).

Scott (1998) documents failed "high modern" efforts in trying to improve the human condition, and points to language itself as the best model for new forms of social life capable of continually adapting broad principles to novel circumstances. Language provides the best available model because it is "a structure of meaning and continuity that is never still and ever open to the improvisations of its speakers" (Scott, 1998, p. 357). Scott (1998, p. 355) proposes an integrity test as a way of interrogating planned, built, and legislated forms of social life. The issues here involve checks on whether the model of language as simultaneously open, malleable, structured, and navigable is adequately approximated. Scott does not, however, remark on the discontinuous multilevel complexity of language, an essential characteristic key to applying it as a model for creating forms of social life with the integrity needed for simultaneously facilitating structured meaning and local improvisations.

Taking language as our model requires information infrastructures sensitive to multiple levels of complexity and the discontinuities that must be negotiated if we are to integrate openness and malleability with structure and navigability (Blok et al., 2020; Bowker, 2015; Bowker et al., 2014, 2015; Marais & Kull, 2016; Star & Ruhleder, 1996). Because "it is not possible simultaneously to perform and objectify an illocutionary act" (Habermas, 1979, p. 43), taking language as a model requires recognizing that:

> no discourse can claim to be free of presuppositions for the simple reason that the conceptual operation by which a region of thought is thematized brings operative concepts into play, which cannot themselves be thematized at the same time. (Ricoeur, 1977, p. 257)

The developmental psychology of complex cognitive and moral operations similarly hypothesizes that "each successive hierarchical integration produces novel understandings by employing the operations of the previous order as conceptual elements in its new constructions" (Dawson, 2003, pp. 335–336). Successive integrations do not replace previous orders but are put into play in a fluid way dependent on other factors, such as the supports available in the environment (Fischer, 1980; Fischer & Farrar, 1987). But specific instances of transitions between levels take place when the unthematized conceptual operations of one stage become the explicit objects of operations at the next stage.

Levels in language correspond with qualitatively distinct denotative, metalinguistic, and metacommunicative statements and with the associated concrete, abstract, and formal levels of cognitive and moral development. Examination of the kinds of meanings shared at each of these levels, and the simultaneity of their usages by persons fluent in a language, suggests how language has already been implemented as a model in the natural sciences and in democratic governance.

That is, broadly and somewhat simplistically speaking, in the natural sciences, data-focused experimental communities of scientists focus, though not exclusively, at the denotative level; instrument-making technicians and standards groups, at the metalinguistic level; and theoreticians, at the metacommunicative level. Analogously, extending the over-simplified model, in democratic governance, the executive branch functions primarily at the managerial level of denotative facts; the legislative branch at the level of collectively projected metalinguistic measures, rules, and laws; and the judicial branch at the constitutional level of metacommunicative theory. Galison (1997) suggests that the disunity of the three communities of scientists revealed in his study of the culture of microphysics makes science stronger and more resilient than it would be if experimentalists, instrument makers, and theoreticians collaborated more closely. Galison (1997, pp. 843–844), like Haraway (1996), finds mechanical metaphors of these groups' interactions dissatisfying and seeks another, more complex image. Galison suggests that image may be found in complex forms of noise-induced order, such as the phenomena of stochastic resonance and associated principles of complex adaptive self-organization, both of which are relevant to the probabilistic form of Rasch's models of measurement (Fisher, 1992b, 2011c, 2017; Fisher & Wilson, 2015).

Preparatory conceptual development is provided by Star and Ruhleder (1996), who suggest that integrating the competing demands for structure and navigability with openness and malleability "create a fascinating design challenge – even a new science." They acknowledge that meeting this challenge will be "highly challenging technically, requiring new forms of computability that are both socially situated and abstract enough to travel across time and space" (Star & Ruhleder, 1996, p. 132). Efforts aimed at meeting those challenges would naturally seem to be allied with the kind of metrologically viable and locally adaptable measurement theory and applications offered in the present volume.

Elucidation of the model of language illuminates ways in which available measurement methods and new variations on them might be further implemented across the sciences, and in expanded conceptualizations of democratic governance and political economy. The separation and balance of powers in this model extends beyond government to the institutional infrastructures of science, law, markets, and communications. Taking language as a model informs concepts of expanded freedom applying across a broad range of contexts, such as employment, education, health care, and environmental resource management, and at different levels of social organization, from individuals to small groups, organizations, and communities to nations and populations. A fuller development of the semiotic theme and how it relates to measurement is needed before a proper account of the integrity test proposed by Scott (1998) can be provided.

6.2.1 Communication in signs

Philosophers characterize language as the vehicle of thought (Derrida, 1976, p. 50; Dewey, 1954, p. 210; Peirce, 1992, p. 30; Wittgenstein, 1958, p. 107) and as providing communications efficiencies akin to an economy of thought (Banks, 2004). Even when thinking takes on a muscular, embodied aspect, or involves the manipulation of nonverbal images, as in the cases of Einstein (1954, pp. 25–26) and Brouwer (1952), sharing of ideas nonetheless demands expression in language. Considered on another level, however, language use is an instance of those kinds of important operations we can perform without thinking, intuitively proceeding without fully understanding the tools we are making use of, which Whitehead (1911, p. 51) referred to as crucial to the advancement of civilization.

Language is the primary medium of both individuals' cognitive processes and their social communications. Speaking in terms of cultural evolution and the dynamics of the processes by which unexamined presuppositions become objects of operations, it is past time to make practical use of the fact that science acts semiotically on models integrating formal ideas, abstract words, and concrete things in systems and does not act directly on things themselves (Cartwright, 1983, p. 129; Brier, 2011, 2013, 2015; Danesi, 2017; Knorr Cetina, 1999; Nersessian, 2008; Sebeok, 2001; Maran, 2007; Nöth, 2018). Practical usage points toward the crucial importance of environmentally embedded infrastructural cognitive scaffolding distinguishing things, words, and ideas in integrated systems. As will be argued at length, today's predominantly ordinal measurement systems in education and other areas exist in a context of infrastructural supports that confuse things and words, taking the linguistic relationships for reality itself. By presuming to describe and access concrete events in ways that ignore the role of language in focusing attention on shared social realities, contemporary conceptual presuppositions conflate levels of complexity and promote dysfunctional automatic associations of numbers and quantities, money and wealth, economic consumption and sexual dominance. This is not a permanently inalterable circumstance. In the same way that the political, scientific, and economic revolutions of the late eighteenth and early nineteenth centuries produced new distributions of semiotically integrated ideals and theories, words and instruments, and things and data, so, too, might today's developments lead to similar possibilities.

The defining and inextinguishable complexity of the situation is such that even nontechnical, everyday language incorporates multiple levels of complexity. That is, the recursive self-reflective nature of language makes it possible for perfectly well-formed statements to be meaningless, as in the classic example of a Cretan who says, "All Cretans are liars," or of a sentence like, "This statement is false." But this paradoxical quality is also the means by which new things come into words. It constitutes the capacity for humor, metaphor, seeing things in new ways, and for learning through what we already know, doing so by making equations between existing knowledge and previously unknown possibilities. And so the abstract way in which a seemingly concrete, denotative statement can paradoxically assert a nonsensical metacommunicative statement about itself is also the way in which a line in a triangle or a number in an equation can implicate another line or number with an irrational length or an infinite number of decimal places.

The value obtained when these paradoxically disjoint and contradictory processes are coordinated at a new level of complexity is central to shifting the paradigm away from alienation in a mechanical clockwork universe toward being at home in a playful participatory universe. Play is the key clue to the methodical aspect of this shift. Just as animals at play metacommunicatively signal to one another that their bites are not intended to hurt (Bateson, 1972, p. 179), so, also, language itself, like natural processes and events, fluidly incorporates multiple levels of meaning and gives hope that humanity can figure out how to give voice to and model its own spontaneously emergent properties as natural in the same ways physical, chemical, and biological processes are. As Gadamer (1989, pp. 104–105) observes, humans play in the same way as animals, as waves on the beach, or as light through the leaves in a breeze. Human play is natural, and as Gadamer suggests, methodically operationalizing playful subjective experience is essential to resituating social life in the context of earthly existence.

Captivation with beauty and desire for meaning set up the play of language games as a semiotic means of extending into new domains the kinds of knowledge infrastructures we take for granted in geometry, science, and everyday language. We typically ignore language's levels of complexity in the design of many information systems and knowledge infrastructures and do so at our own expense (Bowker et al., 2015; Fisher & Wilson, 2015; Rousseau, 1985; Star & Ruhleder, 1996). For instance, in education, health care, and other areas, locally concrete phenomena are not transparently (within a fit for purpose tolerance range of uncertainty) integrated with abstract representations of the formal ideals that are intended to be communicated. Because the existing paradigm's semiotic infrastructure enables a fluid communicability, fallacious conflations of counted events (correct answers, ratings, etc.) are treated as though they are quantities. The transparent illogic involved in routinely interpreting scores of 90% correct as indicative of superior performance has a captivating hold on the public imagination even though everyone is well aware that the nine out of ten questions answered correctly may have been exceedingly easy in relation to the performance one might prefer for a given purpose.

That is, test and survey scores are usually expressed as counts meaningful only within the denotative context of specific questions. They are not typically given as quantities meaningful (within a range of uncertainty) as abstractions applicable across contexts involving different sets of questions. Contrary to widespread assumptions as to the impossibility of deriving such quantities, context-spanning measurements meaningful within stated confidence intervals were shown viable almost 100 years ago (Thurstone, 1926, 1928) and have been in wide, though limited, use in psychology, education, health care, and other areas for decades.

Another way of saying this is to observe that our institutions are not organized around love of beauty as the model of meaning. Though we can acknowledge the distributed effects of collective social processes not under the direct control of any individual or group, systemic biases built into market, governance, educational, healthcare, and environmental institutions nonetheless seem designed with ugly intentions: disempowering and subjugating people (Berti & Simpson, 2021). Efforts at communication are routinely and methodically abandoned in hastily constructed representations obviously and long understood as insufficient to the tasks of fairly mediated relationships. Contrary to popular assumptions, though, we do possess the means we need to address the multiple failures of our cultural institutions to provide access to shared meaning and processes of mutual understanding.

Most diagnoses of these failures take up only the symptoms and not the root causes of the problems. Solutions formulated only within separated, siloed domains of concrete data, abstract instruments, or formal theories, or within particular fields or organizations, must inevitably fail. Even when all three levels of complexity are integrated in a system within a field, community, or department, the resulting plethora of disconnected and incommensurable representations can only ever continue pitting groups in pointless conflicts with one another.

What are needed are theories and methods useful in conceiving, imagining, creating, and improving institutions integrated at metasystematic, paradigmatic, and cross-paradigmatic levels of cognitive and moral complexity (Dawson, 2004; Fischer & Farrar, 1987; Commons & Ross, 2008). Root causal processes integrating bio-neurological, cognitive, and social factors to varying degrees have been successfully identified (Andersson, 2015; Barab & Plucker, 2002; Commons & Duong, 2019; Commons & Goodheart, 2007, 2008; Ekstig, 2010; Latour, 1990, 2005; Commons, Ross, & Bresette, 2011; Ross, 2014; Ross & Commons, 2008; Sutton et al., 2010), but practical paths forward are notably lacking. The diagnoses offered typically and unfortunately omit accessible, integrated combinations of pragmatic methods and deeply resonant meanings essential to effective policy and practice.

Intimidating barriers to such pragmatism are posed by the biological, psychological, and social complexities of individual human beings, complexities that are multiplied several times over by the yet higher order complexities of their interactions and cultural environments. Counterproductive norms supporting or tolerating corruption, inequity, incommensurability, and bias are systemically institutionalized in ways that successfully resist all efforts to change them that are not as complexly structured as the norms they would replace.

The difficulties encountered in figuring out where to start and how to proceed are hard to over-estimate. Fundamental clues, however, suggest the essential importance of beginning from collectively projected patterns of thought and behavior that are always already enmeshed in language's multilevel meaning structures. When these patterns are persistent and stable enough to inform common languages, it may also then be possible to structure meaningful and useful relational integrations of concrete local improvisations, shared abstract representations, and formal explanatory theories in systems. Trial and error experimentation with these relational structures must, notably, be safe and amenable to widespread distribution across societies and social sectors. Most challenging, perhaps, is that multisector legal, accounting, regulatory, scientific, product, and other standards must contextualize – not smother or crush – local negotiations and creative improvisations of meaning.

6.2.2 Semiotically complex measurement

Everyday language and existing metrological standards for scientific units of measurement already satisfy these requirements and provide models on which to build. The arts, especially music, provide clear examples of how beauty and meaning absorb attention in participatory flows of experience that are both uniquely individual and situated within generally communicable technical standards. Creativity in science is not qualitatively different from creativity in art; both are constrained by the quality of the media employed and open onto new possibilities only so far as these are supported by the materials and instruments in use (McLeish, 2019). In the same way that instruments tuned to common scales enable musicians to readily cocreate, so, too, do shared metrics and common currencies expand economic markets by lowering transaction costs (Ashworth, 2004; Barzel, 1982; North, 1990), and so, too, do shared measurement standards and metrological quality assurances expand language's economy of thought (Weitzel, 2004).

Developmentally speaking, the goal is to embody in readily understood tools concepts too technically advanced to be accessible to the general population. Thermometers, for instance, can be read without knowing anything about thermodynamics, and computers can be used without knowing how to fabricate CPUs or how to program. In the same way, we must learn to model, calibrate, and distribute instruments that serve as tools people can use in accomplishing their own self-defined educational, health, employment, social, and other goals. Those tasks cannot be achieved without understanding and strategically situating ourselves in relation to the absorbing powers of beauty and meaning.

A start at accessing such integrations in limited ways has already begun. Domains in which the most obvious institutional failures obscure meaning also simultaneously offer great potential for dramatically enhancing communications and shared understandings. These domains involve the widespread and needless accommodations of obsolete measurement models and methods in education, health care, human resources, organizational performance assessment, environmental management, and many other areas of life. These outdated approaches to quantification needlessly confound and restrict the meaning of counts of correct answers and sums of ratings (survey, questionnaire, test, and assessment scores). These kinds of numeric scores change meaning depending on the particular questions asked, and who answers them. Changing a question or the members of the sample makes the resulting numbers incomparable.

The reductionist assumption that quantitative wholes must be and can only ever be equivalent to the sums of their parts is widely – but mistakenly – recognized and accepted as a fundamental and inescapable epistemological condition. Contrary to popular understandings, mathematicians have held quite a different perspective on mathematics for quite some time. As Chaitin (1994, pp. 12, 13) put it:

> there is randomness in elementary number theory, in the arithmetic of the natural numbers. This is an impenetrable stone wall, it's a worst case. From Gödel we knew that we couldn't get a formal axiomatic system to be complete. We knew we were in trouble, and Turing showed us how basic it was.

There is a profound and troubling contradiction between the fundamental complexities of mathematics and the way mathematics is applied in most educational, health care, human resource management, and environmental policy information systems. The near-universal prevalence of a mathematical blind spot in institutional policy and individual ideation and imagination encompasses a large realm of hidden assumptions. Those assumptions must cease being hidden within unexamined presuppositions and become overt objects of operations if humanity is to succeed in shifting the paradigmatic definition of its institutions – if it is to transition to a paradigmatic stage (Ross, 2008).

Just as children learning a language shift from merely using it in concrete operations ("the cat is on the mat") to remarking on and using its abstract properties ("the word 'cat' cannot scratch"), so, too, must our institutions and sciences also transition from making use of language to acting on it. A qualitative shift in complexity must be initiated away from operating on numbers assumed without evidence or theory to be quantitative wholes equal to the sum of the parts, toward operating on quantities informed by evidence and predicted by theory to be wholes greater than the sums of their parts.

The malevolent and ill-formed seeds of today's institutionalized dysfunctionality continue to sprout, grow, and thrive because they are systematically cultivated and nurtured by unexamined presuppositions hidden within the modern paradigm's metaphors of alienated subjects and objects. Language is indeed the medium of thought, but the *specific* terms of this condition of dependency are not as irrevocable as is usually assumed. Bounded rationality is not inherently or permanently imprisoned within a given language's conceptual limits. The question is how ecologically complex institutional infrastructures might be organized in ways that compactly represent mathematically and linguistically complex aspects of the real world in accessible and meaningful ways to populations unable to create them for themselves.

6.2.3 An integrity test for proposed improvements to the human condition

Scott (1998, p. 355) proposes an integrity test aimed at informing evaluations of efforts intended to improve the human condition. His historical account of repeated failures in this regard leads to the formulation of three questions:

1. To what degree does the new form of social life promise to enhance the skills, knowledge, and responsibility of those who are a part of it?
2. How deeply is the new form of social life marked by the values and experiences of those who compose it?
3. Is it possible to distinguish rigidly imposed categories of social life that permit little or no modification from situations that are largely open to the development and application of local, polyvalent, practical know-how?

Moss (2004, p. 224) cites Scott (1998) when warning about the risks of heavy-handed top-down impositions of too-rigid definitions of coherence between classroom assessment and accountability standards (Wilson, 2004). But coherent measurement systems need not necessarily be defined in such a reductionist way, as is evident in hierarchically complex applications of probabilistic models for measurement (Commons et al., 2008, 2014; Dawson, 2003, 2004; Dawson-Tunik et al., 2005; Mueller et al., 1999; Mueller & Overton, 2010; Overton, 1998). Alternative answers to Scott's questions can be formed in light of his recommendation that language be taken as a model of structures as open and malleable as they are continuous and navigable.

Applications of formative assessments in education inform an answer to Scott's first question focused entirely on the purpose of enhancing the skills, knowledge, and responsibility of the participating students and teachers. Meta-analyses of hundreds of studies of formative feedback's effects on outcomes provide extensive evidence of enhanced skills, knowledge, and responsibility among those using it (Black & Wiliam, 1998; Hattie, 2008). Given the adage that we manage what we measure, parallel applications in human resource, citizen participation, and other messaging systems in a wide range of organizations can be imagined.

Similarly, regarding the second part of Scott's integrity test, by incorporating students' and teachers' questions and answers, formative assessment messaging systems are forms of social life that embody the values and experiences of those who compose it. The denotative individual student responses, the metalinguistic empirical structural invariances, and the metacommunicative explanatory theories all emerge from and are evaluated relative to nothing but the values and experiences of the students answering questions posed to them by their teachers and others in the educational community. Analogous messaging systems in health care, social services, environmental management, and other areas could certainly be designed to also represent their participants' values and experiences.

Finally, taking up Scott's third question, rigid impositions of categories of social life that permit little or no modification are certainly common in many efforts intended to facilitate coherent messaging in education. This was the case with the No Child Left Behind legislation, as documented by Ladd (2017). In addition, the need for a particular kind of workforce, or the desire to exclude taboo subjects from the curriculum, may give rise to the kind of contrived theories and facts that historically have been used to justify power relations.

But Scott's question is whether these contrivances will be distinguishable from "situations that are largely open to the development and application of local, polyvalent, practical know-how." And the answer to this question is that, yes, clear distinctions of these different kinds of situations ought to be evident in the balance of powers effected by the clear separation of executive, legislative, and judicial functions at the denotative, metalinguistic, and metacommunicative levels. Most significantly, at the local level of individuals, students and teachers will see when assessment responses sensibly approximate or depart radically from the expected learning progression, and when the questions asked in an assessment pertain to the relevant learning objective (Chien, et al., 2018; Linacre, 1997; Wright et al., 1980). Students who care about learning, and teachers who care about students, will know where they stand relative to where they were yesterday, where they want to be, where everyone else is, where their special strengths and weakenesses, if any, lie, and what to do next, as shown in existing navigations of learning progressions (Black et al., 2011; Fisher, 2013; Dozier et al., 2023).

Coherent information infrastructures intentionally built to span discontinuities in complexity ought then to provide new forms of previously unavailable leverage connecting them with accountability expectations (Fisher, 2023a; Fisher, Oon, & Benson, 2021; Fisher & Wilson, 2015). As suggested by Star and Ruhleder (1996), we need to shift the focus of authority and responsibility in education for individual student outcomes away from homogenized top-down controls and standards heedless of language's systemic discontinuities. The spontaneous bottom-up self-organization of locally directed outcomes coherently connected across levels of complexity seems likely to embody features of past successes in human social history, such as the innovative productivity of science's discontinuous communities of theoreticians, experimentalists, and instrument makers (Galison, 1997; Galison & Stump, 1996) and the separation and balance of powers in government's executive, legislative, and judicial branches. Though the creation of such multilevel systems will take careful planning and development, identifying the problem may turn out to have been the most difficult task of all.

6.3 Augmenting logic with aesthetics and ethics

6.3.1 Locating the limits of logic and method

Various articulations of the logic and methods of measurement demonstrated in the chapters of this book have been available and in use for over 50 years (Bradley & Terry, 1952; Brink, 1972; Fischer, 1968, 1973; Guttman, 1944, 1950; Rasch, 1960, 1961; Loevinger, 1965; Luce, 1959; Luce & Tukey, 1964; Wright, 1965, 1968), and in some respects, for almost 100 years (Thurstone, 1926, 1928; see Andrich, 1978, 1988, p. 43; Brogden, 1977). Over the course of those years, the inferential advantages and meaningful comparability provided by this logic and these methods (Andrich, 2010; Embretson, 1996; Rasch, 1977; Narens & Luce, 1986; Wilson, 2005, 2013b; Wright, 1977, 1997; Wright & Masters, 1982; Wright & Stone, 1979) led to their adoption across wide ranges of research and practice (Alagumalai et al., 2005) involving everything from educational (Confrey et al., 2021; Connolly et al., 1971/2007; Green, 1986; Masters & Keeves, 1999; Wilson, 2018; Wright, 1977, 1997), psychological (Commons et al., 2014; Dawson, 2004; K. Fischer & Dawson, 2002), and health care outcomes (Allen & Pak, 2023; Belvedere & de Morton 2010; Bezruczko, 2005; Cano et al., 2009; Christensen et al., 2013; A. Fisher, 1997; Massof & Bradley, 2023) to clinical chemistry (Fisher & Burton, 2010), climate-neutral environmental policies (Fisher, Melin, & Möller, 2021), human resources (Barney & Fisher, 2016), sociology (Duncan, 1984a/b/c), and the psychophysics of vision (Brown et al., 2014; Karakus et al., 2018; Massof et al., 2024; Powers & Fisher, 2021), and to measurements of the outcomes of spiritual care (Snowden et al., 2022; Pugliese et al., 1993) and mindfulness practices (Medvedev et al., 2016; Solloway & Fisher, 2007). Especially intensive applications have emerged in high stakes contexts demanding legally and scientifically defensible results, such as graduation, admissions, accountability, licensure, and certification decisions (Bergstrom & Lunz 1999; Grosse & Wright 1986; Han et al., 2022; Kelley & Schumacher 1984; Li et al., 2021; Masters, 2007; Nungester et al., 1991; O'Neill et al., 2005; Smith et al., 1994; Tsai et al., 2013; Wendt & Tatum, 2005; Zhang & Roberts, 2012).

In addition, dating from the earliest publications in this area, the inferential equivalence across physics and psychology of the quality of the quantitative results produced by this logic and these methods has been repeatedly explained and demonstrated (Andrich, 2004, 2005; Fisher, 2003, 2010d; Fisher & Stenner, 2013; Luce & Tukey, 1964; Narens & Luce, 1986; Rasch, 1960, pp. 110–115; Thurstone, 1928; Wilson, 2013a; Wright, 1997). As Rasch (1968, p. 26) put it,

> Whether observations come from Physics, Psychology, Social Sciences or Humanities, and whether they be quantitative or qualitative, gives no apriori reason for believing in or abolishing methods founded upon strong probabilistic models.

Ongoing support for the validity of these claims by cross-disciplinary teams of metrological and psychometric collaborators (Cano et al., 2019; Mari & Wilson, 2014; Mari et al., 2023; Pendrill, 2014, 2019; Pendrill & Fisher, 2015; Pendrill & Melin, 2023) makes it natural to wonder why this logic and these methods have not been much more widely adopted in psychology and the social sciences.

To put the matter directly: Given the availability of forthright logic and evidence supporting a transformed theory and practice of measurement, why is it the case that, not only has an expanded SI not emerged, why are the pros and cons of such an idea not even discussed? The very concept seems, in the mainstream conception, so far-fetched as to be laughable, or to make its serious consideration to be a cause of fear and trepidation. But perhaps we ought to pause a moment and reflect, in the spirit of Eleanor Roosevelt's (1960/2011, pp. 29–30) counsel that the thing we fear is what we must do. We might also adopt the likeminded spirit of Maurice Sendak's (1963) monstrous wild things that turn out only to want to play when you look them in the eye.

In taking that pause, we need to ask: What are the consequences of continuing to accept as valid epistemologies we know to be wrong, as Bateson (1972, p. 493) observed we commonly do? For example, how can it be that everyone knows your two rocks might be any amount more or less rock than my eight rocks, but we not only willingly accept counts and percentages of correct answers and ratings as supposed measurements (Bateson, 1978; Thurstone, 1937, p. 231; Wright, 1994), but go so far as to embed them in the institutional infrastructures of education, markets, health care, social services, government, management, etc.? We would seem to allow this illogic to continue only because of the widespread belief that there is no viable alternative. But given decades of research and practice demonstrating clear alternatives, one has to wonder if something more is going on.

Do the modern metaphysics of "strong objectivity" and the descriptive role of science in relation to a supposedly independently existing universe amount to anything more than a pathological excuse allowing us to avoid taking responsibility for our creations? Is not continuing to act falsely – as though more meaningful measurement is optional, as though it is a preference that can be chosen or rejected depending on uncontrollable, fateful combinations of data quality and construct definitions – tantamount to admitting current practices constitute fraudulent malpractice (Grimby et al., 2012)? Given that (a) the laws of nature demand expression in language and technology to be communicated, demonstrated, and used and (b) that language and technology are plainly and simply products of human interests, can the positivist realism of a nature existing entirely independent of human interests be anything but a sham manipulation by the powers that be having no other purpose apart from maintaining the status quo? Do not the ideals of justice and equity inspire us to do better (Elnegahy et al., 2022; Mallinson, 2024; Mallinson et al., 2022; McNamara et al., 2019; Mtsatse & Combrinck, 2018; Russell, 2023; Sul, 2024)?

The laws of science, of course, may well be completely uniform throughout the entire universe, but practical applications of those laws absolutely require human-made technologies. And, conversely, postmodern cultural and historical relativizations of those laws similarly also cannot be communicated or accessed in practice except by means of linguistic standards. Here, largely unrecognized by both modern and postmodern perspectives, humanity confronts a fundamental and systemic dissonance in its thinking. Both modern transcendentalism and postmodern relativism can proceed only in terms of distributed networks of technical standards, as was well understood by Derrida (2003, p. 63) when he acknowledged that his deconstructive process advances "always by measuring the distance from the standards I know or that I've been rigorously trained in." The seeming opposition of modern and postmodern as bitter enemies conceals their deeper common assumption of a modern metaphysics ignoring the ways that standards mediate ideals and local circumstances (Latour, 1990, 1993). As Latour (1991, p. 17) put it, "Postmodernism is a disappointed form of modernism. It shares with its enemy all its features but hope."

By systematically ignoring the roles of linguistic and technical standards in relation to ideas and things, and to theories and data, humanity has put itself in a classic double bind, the kind of irreconcilable internal inconsistency that has long been characterized as schizophrenic (Bateson et al., 1956; Bateson, 1972, pp. 206–217; Bateson, 1991, p. 203; Boundas, 2018; Deleuze & Guattari, 1977; Star & Ruhleder, 1996; Fisher, 2021). In a manner akin to a parent demanding that a child stop acting like a child, rendering the child childish whether they obey or disobey, we paradoxically

- accept simplistic forms of psychological and social measurement as a necessary evil at the same time that we say – against the evidence – that measurement is inherently reductive and impossible to achieve in psychology and the social sciences in any form even remotely comparable to its forms and functions in the natural sciences;
- deplore measurements that numerically reduce and homogenize complex humanity, but do so in language applying the same symbolic logic and inferential processes as are used in mathematical language;
- project unnecessary, reductionist consequences on social and psychological quantification despite making use of heterogeneously interpreted irreducible applications of physical measurements dozens of times a day; and
- in so doing, narcissistically impose on measurement theory and practice the very same reductionist oversimplifications measurement is falsely accused of necessitating.

Our schizophrenia is one in which we accept the necessity of high-stakes roles for psychological and social measurement while we also feign an increasingly willful ignorance that knowingly, systematically, fraudulently, and cynically presents numeric counts as measured quantities. Hundreds of times a day, virtually everyone among us uses language's standardized phonemic, alphabetic, numeric, and grammatic abstractions to negotiate shared understandings of associations between unrealistic formal ideals and unique concrete local circumstances, while we also refuse to recognize, accept, and adopt ways of doing this involving social and psychological measurements.

In the same vein, we willingly invest several times more resources in metrology and standards than in research as a whole, with impressive returns (NIST, 1996, 2009), but the evolution of standards networks associated with growing sciences (Latour, 1987, 1993, 2005) is systematically ignored in favor of a magical ideation asserting that those standards exist in nature independent of human interests, and in favor of perversely and counterproductively persisting in *not* interpreting repeatably reproduced and predicted psychological and social unit quantities as evidence justifying decisive action initiating new SI standards networks.

In this state of abject denial, as suggested by Latour (2012), do we not perpetually reproduce Dr. Frankenstein's fear-induced abandonment of his creation and his subsequent blaming of it for monstrous acts that in actual fact the Dr. himself was responsible for? Can we really justify continuing to act as though nature plays its roles in our lives in exactly the same ways whether or not we provide technical media and institutional contexts for its manifestations and their incorporation into our routines and habits? How long can we sustain this unsustainable schizophrenic disconnect between our self-image as just and equitable and our unjust and inequitable lived realities? What will it take to crystalize thought and action focused on creating new knowledge infrastructures extending the concrete, abstract, and formal levels of complexity in language into new domains? Is it possible to inspire new self-organized movements focused on embedding mathematical equivalents of the scales of justice deeply within the messaging systems of our governance, market, legal, educational, and communications institutions?

The salience of the question as to why advanced measurement is not more widely implemented can be amplified at the same time we strategically address anxiety about doing so, by pointing out:

- that a de facto measurement standard for English reading comprehension productively integrating assessment and instruction for over 30 million students per year in over 80 countries has been commercially available since 1995 (Stenner et al., 2013), with analogous standards for Spanish and Arabic also available;
- that a similar theory-based approach to validating the Common European Frame of Reference for measuring and managing language proficiency has emerged over the last 30 years as a medium for integrating assessment and instruction (North, 1995, 2014, 2020); and
- that a new consensus standard for the metrology of short-term memory and attention span has been in development in a multinational European project for the last several years (Cano et al., 2018, 2019; Melin et al., 2023; Quaglia et al., 2016, 2019).

Plausible answers to the question as to why advanced measurement and metrology have not become more widely adopted can be discerned in the general discourse of reflections on social and psychological measurement. Misconceptions abound in this area; only a small fraction of publications make use of even the most longstanding and well-established quantitative theory and practice focused on invariant quantities and uncertainties.

As was pointed out by Latour (1990), this oversight, this failure to learn the language of the field of interest, is especially unexpected in science and technology studies. This area in the social studies of science was initially referred to as "laboratory ethnography." Ethnography, of course, involves the basic methodology of participant observation, which typically requires learning the language of the culture involved and becoming intimately familiar with its practices. But Latour's (1990, p. 146) remark on this situation, made soon after the emergence of the field of science and technology studies, remains today as relevant as it was then, over 30 years ago:

> the few people, myself included, who have used ethnographic methods to get at modern sciences have used the most outdated version of anthropology: the *outside* observer who does not know the language and the customs of the natives, who stays for a long time in one place and tries to make sense of what they do and think by using a metalanguage which is as distant as possible from those of the natives who are not supposed to read what he writes. As Woolgar has pointed out many times, this is a very naive version of the naive observer–a version that is now abandoned in mainstream ethnography and which seems to survive only in so called "lab studies."

Mathematics is widely recognized as the language of science and nowhere is mathematical modeling of greater importance than in the theory and practice of measurement. Mathematically sophisticated philosophical and anthropological investigations – ones focused specifically on scientific and not merely statistical models of measurement – are nonetheless exceedingly rare.

Of course, as has long been recognized in psychological measurement, mathematical understanding need not always take the form of a capacity for dealing with equations and numbers. Thurstone (1937, p. 231) not only distinguished between counts and quantities, he recognized mathematics as the language in which science thinks, and that students lacking skill in calculation may nonetheless be fertile with experimental ideas, while others with great facility in manipulating equations may lack the flexibility of mind essential to creative science. Guttman (1994) similarly noted the qualitative origins of mathematical comparisons in everyday language and Husserl (1970a/b) founded his phenomenological philosophy on mathematical concerns without stating any equations. Heidegger (1967; Harries, 2010; Kisiel, 2002; Roubach, 2008) gave an intensive and insightful analysis of the ancient Greek category of the mathematical in entirely qualitative terms, expanding on the root concepts of *mathesis*, learning, and *ta mathemata* as learning through what is already known, and what is learned in that way. Unfortunately, these issues are only rarely related to contemporary measurement modeling (Fisher, 1988, 1992, 2003, 2004, 2010b, 2023).

Discussions and examples of culturally specific measurements are given by Mtsatse and Combrinck (2018), Mallinson (2024) and Sul (2024). They provide particularly relevant examples of the challenges involved in following Latour's (2012) recommendation that we learn to care for our technologies as we do our children, instead of abandoning them to their fates and allowing them to become monstrous. Learning the languages of measurement technologies will just be the beginning of coming to understand how to conceive, gestate, midwife, and nurture them to maturity and fulfilling ends. To succeed we will need to grasp and try to live up to the ethical importance of pregnancy as the site of reliant maternal eroticism's mother love capacity for providing access to language (Kristeva, 1980, 2014).

In making use of concepts and tools we do not create alone and cannot fully control, in accepting beauty as the model of meaning, in giving ourselves over to the play of language games, in learning how to systematically embody new concepts and things in words distributed throughout sociocognitive ecosystems, and in accepting the gift of life, we must, as Latour (2012) puts it, care for our technologies as we do our children. In so doing, we take the "ontopoiesis of life as a new philosophical paradigm," and

> We move away from the classical prejudice that mathematics involves "calculability" only, in a qualitative, aesthetic expansion of the discipline. The abstract science of mathematics "humanizes" itself. (Tymieniecka, 1998, p. 23)

Much, almost everything, remains to be done to put this paradigm shift and cultural transformation in play, but it may turn out that identifying the problem was the primary obstacle to be overcome.

Discussions of measurement's social value or lack thereof typically involve misconceptions of oversimplified, autonomous individuals, and positivist realism. After summarizing the arguments involved in these hypothetical explanations of the failure to widely adopt measurement models and methods that are more rigorously defined, meaningful, and useful than is typically the case, an alternative explanation is proposed. This alternative addresses the need for complementing the logic of measurement with a fuller absorption of subjective experience into an aesthetics and ethics satisfying the desire for beauty and justice.

6.3.2 Three misconceptions of quantity and quantification

The basic point is an expansion on the observation (Dear, 1992; Porter, 1995, 1999; Wright, 1958) that objectivity has been too narrowly conceived as completely excluding subjectivity. As Galison (2008, p. 293) put it, objectivity and subjectivity mutually implicate one another, and so we "need a joint epistemic project addressing the historically changing and mutually conditioning relation of 'inside' and 'outside' knowledge." An entirely different paradigm follows from seeing subject and object paradoxically unified yet separable in distributed multilevel networks of words and instruments traceable to linguistic and metrological standards (Fisher, 2019, 2020, 2023a). Furthermore, instead of apolitically by-passing the social, and instead of considering only a broadly generalized conception of social concerns as co-produced and co-evolving with measurement, the need for alignments of property rights, financial markets, and communications networks with metrologically quality assured quantities is emphasized.

As Overwijk (2021, p. 144) puts it, "the 'operational closure' of sociotechnical systems of measurement . . . in fact produces the historical logics of technical reason and, paradoxically, also generates spaces of critical-political openness." The technical details of how spaces for this openness are cleared and inhabited have the satisfying quality of being variations on familiar ways of clearing paths forward, as is illustrated via references to the other chapters in this book.

First, quantification is widely misconceived as inherently, inevitably, and unequivocally reducing rich complexity to manageable numbers, with the aim of making individuals and societies more easily controlled (Merry, 2016, 2019; Merry & Wood, 2015; Postman, 1992; Porter, 1999; Power, 2004). Though any language can be used in this way, the historical uses of mathematical language in quantifying human attributes have confused number and quantity, with the highly deleterious effects of homogenizing, erasing, and ignoring important differences. Everyday language and thinking are already thoroughly mathematical before any numbers or equations become involved. That rigorous quantification does not necessitate reductive oversimplification comprises a fundamental challenge that will have to be met for improved outcomes communications and management to be realized.

Second, that challenge is rendered yet more intimidating by another double bind. Even when they address matters of unified collective consciousness, logical arguments intended to educate and persuade individuals serve only to play into the perpetuation of Western dualisms. By failing to manifest, concretize, estimate, and distribute collectively constituted constructs, we can only default to the existing predominant modern cultural presumptions that individual thinking is rational in itself and that it is autonomously controlled within our brains. The deeply ingrained assumption of self-sufficient individuality is a primary obstacle to fulfilling the potential of advanced measurement and metrology. Articulating and operationalizing an alternative conception of quantification must instantiate a new paradigm of bounded rationality (Kahneman, 2003; Foxon, 2006), where ecologies of mind and infrastructure take the individual in its environment as the unit of survival (Bateson, 1972, 1978; K. Fischer & Farrar, 1987; R. Fisher, 1934a; Lerner & Overton, 2017; Meloni, 2019; Noble, 2017; Watts, 1973).

What we need are metrological infrastructures incorporated in well-designed and resourced systems of entrepreneurship: “resource allocation systems that combine institutions and human agency into an interdependent system of complementarities” (Acs et al., 2018, p. 501). This theme capitalizes on the ways that knowledge infrastructures, such as standardized phonemes, alphabets, fonts, grammars, unit quantities, and protocols, provide supports embedded in the external environment enabling more or less advanced thinking on the parts of individuals. In recognizing that “the fundamental concepts of measurement can be extended to embrace any homomorphic representation by a symbol system” (Finkelstein, 1975, p. 223) we see how to act on the fact that “cultural progress is the result of developmental level of support” (Commons & Goodheart, 2008), and “that organism and environment are inseparable in cognitive development” (K. Fischer & Farrar, 1987, p. 646). As Dewey (1954, p. 210) wrote

> Meanings run in channels formed by instrumentalities of which, in the end, language, the vehicle of thought as well as of communication, is the most important. A mechanic can discourse of

> ohms and amperes as Sir Isaac Newton could not in his day. Many a man who has tinkered with radios can judge of things which Faraday did not dream of.

In short, today's cognitive models focused on mental processes occurring within individual brains do not operationalize the conceptual instrumentalities we need to channel new flows of meaning. Modeling complex wholes greater than the sums of their parts will instead have to tap into knowledge infrastructures making available collective projections of invariantly structured constructs not under the control of any individuals or groups.

The paradox is one of figuring out how to set up institutional infrastructures that distribute new capacities for everyone to "intentionally not intend," in the same manner that jazz musicians similarly must learn to let go of conscious control (C. Fisher et al., 2021) and allow an embodied understanding to have its way. Dancers and athletes likewise develop advanced coordinations of physical capacities that are not consciously controlled. Effective analogies from these kinds of embodied understandings are found in everyday linguistic fluency (Abram, 1996), and the over-learning of instruments tuned to common scales used in playing standards accepted as normative within the social environment and made accessible in transdisciplinary contexts.

Third, cultural expectations infusing modern Western values lead to the general assumption that quality assured metrological unit systems for the measurement of manufactured capital have been spontaneously propagated for free as an automatic consequence of their supposed objectively real existence, despite the wealth of documented evidence to the contrary (Alder, 2002; Bernstein, 2004; Jasanoff, 2004, 2015; Latour, 1993; Miller & O'Leary, 2007; Mirowski, 2004; Schaffer, 1992; Shapin, 1989). Concomitantly, it is assumed that the lack of a similar magical spontaneous propagation of metrological units for the various forms of human, social, and natural capital being measured follows irrevocably from their status as subjective expressions of human interests.

But it is historical fact that common languages for measuring and managing the scientific, financial, legal, and communications aspects of manufactured capital were crucial to the scientific, economic, and democratic revolutions of the eighteenth and nineteenth centuries (Alder, 2002; Bernstein, 2004; De Soto, 2000). Though the historical processes by which these common languages were created and continue to evolve tend to be ignored and devalued, the pertinent lesson to be learned from them concerns how their successes might be replicated via targeted investments in creating efficient market institutions aligning human, social, and natural capital with financial profits (Fisher, 2002, 2010a/c, 2011a, 2012a/b, 2020, 2023b).

Coming at matters from a different point of view, Cliff (1993, p. 87) offers another reason why "the interval-scale status" of psychological and social "variables is not compelling." His explanation concerns "the amount of error involved." Of course, if this reference to "the amount of error" concerns the lack of contractual guarantees as to a defined SI unit's quality-assured traceability to a fit-for-purpose realization, Cliff is completely correct in his assessment. The typical approach to interval scale estimation in psychology and the social sciences is completely ad hoc and even then does not concern itself with attempting to reproduce the quantities established in prior published applications of an instrument. So, if it should turn out that multiple varying logit calibrations of the same instrument should prove to be linearly related, determining that from existing publications indeed involves so many sources of uncontrolled error that the interval scale status of the variable would not be a compelling factor in arguments for improved measurement.

It may be, however, that Cliff is ignoring the expensive, meticulous, legally defensible, financially accountable, and ubiquitous quality assurance efforts invested in ensuring that the uncertainties associated with interval units informing physical measurements are minimized to fit-for-purpose tolerances (Pendrill, 2006, 2019; Pennecchi et al., 2022; Weitzel & Johnson, 2012). The systematic erasure of the microprocesses that make measurement results comparable outside laboratory walls (Ackermann, 1985; Latour & Woolgar, 1979; Schaffer, 1992; Shapin, 1989) has created a metaphysical realm in which unit quantities are assumed to somehow magically propagate across time and space for free. It is quite as though the enormous sums invested in and returned as profits from metrological processes and systems (Gallaher et al., 2007; Latour, 1987; NIST, 1996, 2009; Poposki et al., 2009; Seiler, 1999; Semerjian & Watters, 2000) are somehow extraneous factors unrelated to producing the desired results. But nothing could be further from the truth.

That said, per Cliff's remark, the uncertainties (standard errors) tolerable across psychological and social measurement applications vary, of course, with the demands for precision, such that screening tools need only distinguish two groups, at a traditional reliability of 0.67 or so, to be fit for purpose; while diagnostic tools need to distinguish three to five groups, with reliabilities of 0.80–0.94; high stakes and accountability applications need the legally defensible precision of separating six or more groups (0.95 reliability coefficients); and research on small effect sizes may need even smaller tolerances (0.97 and higher). Research on well-designed and properly administered instruments shows the dependable stability of constructs over samples and item selections, where invariances are repeatedly reproduced to the intended degree of precision (Elbaum et al., 2011; Fisher, 1997a/b, 1999; Fisher et al., 1995a/b, 1997, 2012; He & Kingsbury, 2016; Morrison & Fisher, 2021; Stenner et al., 2013). Implementations of an expanded SI will require carefully designed and maintained assurances leveraging these fit-for-purpose capacities.

Uncertainty is a matter of central concern in metrology (JCGM, 2008), though end users are usually afforded the convenience of ignoring it by instruments calibrated to the needed precision. The engineering of quality assured fit for purpose tolerances and interoperability is essential in commercial production contracting (Pendrill, 2019; Pennecchi et al., 2022; Weitzel & Johnson, 2012). Echoing Duncan's (1984b, pp. 38–39) call for social measurement to be brought within the scope of historical metrology, analogous needs were identified and posed in sociocognitive measurement research almost 25 years ago:

> The task of psychosocial measurement has another aspect that remains virtually unaddressed, and that is the social dimension of metrology, the networks of technicians and scientists who monitor the repeatability and reproducibility of measures across instruments, users, samples, laboratories, applications, etc. For the problem of valid, reliable interval measurement to be solved, within-laboratory results must be shared and communicated between laboratories, with the aim of coining a common currency for the exchange of quantitative value. Instrument calibration (intralaboratory repeatability or ruggedness) studies and metrological (interlaboratory reproducibility) studies must be integrated in a systematic approach to accomplishing the task of developing valid, reliable interval measurement. (Fisher, 2000, p. 529)

Developments in the field over the years since this passage was written substantiate the viability of the metrological vision. Not only will new standards of practice and regulatory criteria need to be determined for applications of an extended SI, those utilitarian and technical issues will also need to be fit into a transformed world hypothesis, one with a participatory logic, aesthetics, and ethics attracting voluminous and intensive investments in creating clear and preferable alternatives to the modernist root metaphor of an independently existing and alienating clockwork universe. Though this transition has been envisioned in many different variations (Capra & Luisi, 2014; Cole, 1985; Kauffmann, 1996; Laszlo, 2019; Sahlins, 2022; Swimme & Tucker, 2011; Wheeler, 1994; Zukav, 1979), pragmatic paths forward in an operational program have been lacking. It may be that the globally manifest pent-up demand for change will soon find its way toward articulating, infrastructuring, and enacting new kinds of ecological political economies.

6.4 Participant observers across the sciences

Though his characterization of a “postmodern science” would be better expressed in unmodern or amodern (Dewey, 2012; Latour, 1990, 1993) terms, Toulmin (1982, p. 97) sets the stage for a method of measured participant observation applicable across the sciences saying

> As we now realize, the interaction between scientists and their objects of study is always a two-way affair. There is no way in which scientists can continue to reduce the effects of their observations on those objects without limit. Even in fundamental physics, for instance, the fact that subatomic particles are under observation will make the influence of the physicists’ instruments a significant element in the phenomena themselves. As a result, during the twentieth century scientists have had to change their interpretive standpoint not merely in the human sciences but elsewhere. In quantum mechanics as in psychiatry, in ecology as much as in anthropology, the scientific observer is now – willy-nilly – also a *participant*. The scientists of the mid-twentieth century, then, have entered the period of postmodern science. For natural scientists today, the classical posture of pure spectator is no longer available even on the level of pure theory; and the objectivity of scientific knowledge can no longer rely on the passivity of the scientists’ objects of knowledge alone. In the physical sciences, objectivity can now be achieved only in the way it is in the human sciences: the scientist must acknowledge and discount his own reactions to and influence on that which he seeks to understand.

The physics Nobelist Wheeler (1974, p. 689) implicitly expands on Toulmin's point saying

> One view holds that as we keep on investigating matter, we will work down from crystals to molecules, from molecules to atoms, from atoms to particles, from particles to quarks – and mine to forever greater depths. A very different concept might be called the "Leibniz logic loop." According to this view the analysis of the physical world, pursued to sufficient depth, will lead back in some now-hidden way to man himself, to conscious mind, tied unexpectedly through the very acts of observation and participation to partnership in the foundation of the universe. To write off the power of observation and reason to make headway with this question would seem to fly against experience.
>
> I see no more hopeful sign that we can and will make our way into this unknown land than the immense progress we have already made into the world of the quantum, where the observer and the observed turned out to have a tight and totally unexpected linkage. The quantum principle has demolished the once-held view that the universe sits safely "out there," that we can observe what goes on in it from behind a foot-thick slab of plate glass without ourselves being involved in what goes on. We have learned that to observe even so miniscule an object as an electron we have to shatter that slab of glass. We have to reach out and insert a measuring device. We can install a device to measure position or insert a device to measure momentum; but the installation of the one prevents the insertion of the other. We ourselves have to decide which it is that we will do. Whichever it is, it has an unpredictable effect on the future of that electron, and to that degree the future of the universe is changed. We changed it. We have to cross out that old word "observer" and replace it by the new word "participator." In some strange sense the quantum principle tells us that we are dealing with a participatory universe.

Wheeler raises three questions in this participatory context helpful in setting out directions for future research. The answers pull together themes from different fields with the pragmatic aim of informing complex integrations of seeming opposites in ways that lead to a diverse array of somewhat convergent and somewhat divergent actionable programs for metrological research and development.

6.4.1 The mystery of the mind: "Consciousness can analyze the world around; but when will consciousness understand consciousness?"

Wheeler's first question, developmentally speaking, points toward the ways in which understanding in general is obtained when unexamined assumptions informing operations from a hidden, subconscious level are explicated and are themselves made into objects of operations (Commons & Richards, 2002; Dawson-Tunik et al., 2005; K. Fischer, 1980; K. Fischer & Farrar, 1987). This process, of course, entails the emergence of another sphere of unexamined assumptions which may also be explicated in a potentially infinite sequence (Bateson, 1972, p. 183; Star & Ruhleder, 1996, pp. 117–118) of such transformations. Focusing attention for a moment on a basic transition in this sequence (Bateson, 1972, p. 183; Star & Ruhleder, 1996, pp. 117–118), a child able to say "the cat is on the mat" may be unaware of using language. Sometime later, the same child may say "the word 'cat' cannot scratch," humorously marking the transition from a concretely denotative factual observation of something learned to an abstract metalinguistic representation of learning about learning. Past that, the child may yet later transition from a statement about a word to a metacommunicative statement about a statement, such as "my telling you where to find your cat was friendly." Now the child has complemented concrete facts and abstract words with a formal theory of learning.

Understanding consciousness then entails explicating the multilevel characteristics of its embodied experience and the extension of that physicality in language. Just as organic embodiment entails involuntary metabolic functions not under conscious control, so also does embodied consciousness also entail an infinite depth of cognitive functions operating subconsciously (Abram, 1996; Chernero, 2013; Harding & Hintikka, 2003; Hotton & Yoshimi, 2010; Ihde, 2002; Irigaray, 1984; Lakoff & Johnson, 1980, 1999; Lakoff & Núñez, 2000; Latour, 2004; Merleau-Ponty, 1964; Olteanu, 2021; Overton, 1994a/b, 1997, 2008; Polanyi, 1974). No one has ever comprehensively understood the reasons why any language's phonemes, grammar, syntax, word forms, etc. are structured in the arbitrary ways they are; this lack of conscious understanding does not, however, compromise language use.

Written language and technology extend spoken language's embodied understanding in ways that also advance civilization by making it possible to execute operations we do not understand, as Whitehead (1911, p. 61) put it. Science is a primary means by which embodied understanding is advanced, as is evident in there being no need to understand thermodynamics in using a thermometer. This is also evident in the way that, historically, theoretical explanations of phenomena follow from access to technologies allowing experimental play with controlled effects (de Solla Price, 1986; Hankins & Silverman, 1999; Ihde, 1983; Kuhn, 1977; Latour, 1987, 2005; Nersessian, 1996). De Solla Price (1986, p. 240) remarks that:

> Historically, we have almost no examples of an increase in understanding being applied to make new advances in technical competence, but we have many cases of advances in technology being puzzled out by theoreticians and resulting in the advancement of knowledge. It is not just a clever historical aphorism, but a general truth, that "thermodynamics owes much more to the steam engine than ever the steam engine owed to thermodynamics."
>
> Historically the arrow of causality is largely from the technology to the science.

Kuhn (1977, p. 90) concurs saying

> Of the nine pioneers who succeeded, partially or completely, in quantifying [energy] conversion processes, all but Mayer and Helmholtz were either trained as engineers or were working directly on engines when they made their contributions to energy conservation. Of the six who computed independent values of the conversion coefficient, all but Mayer were concerned with engines either in fact or by training. To make the computation they needed the concept [of] work, and the source of that concept was principally the engineering tradition.

Science effectively extends embodied language by learning how to conceive ideas representing physical experience. Playful absorption into dialogues with a persistently identifiable effect allows the articulation of a coherent account of its properties and constitutes the methodical process by which the mathematical object becomes something learned through what is already known. This is the point at which consciousness begins to understand itself.

Heidegger refers to this process when he holds that, "upon the basis of the mathematical, *experientia* becomes the modern experiment." He implicitly spells out the need for an extended SI writing

> Because the [mathematical] project establishes a uniformity of all bodies according to relations of space, time, and motion, it also makes possible and requires a universal uniform measure as an essential determinant of things, i.e., numerical measurement. The mathematical project of Newtonian bodies leads to the development of a certain "mathematics" in the narrow sense. The new form of modern science did not arise because mathematics became an essential determinant. Rather, that mathematics, and a particular kind of mathematics, could come into play and had to come into play is a consequence of the mathematical project. (Heidegger, 1967, p. 93)

We can restate this by noting how repeated demonstrations of structural invariances across samples and instruments in social and psychological measurement establish uniformity in social forms of life according to embodied developmental relations. These relations integrate individuals' thought processes with cognitive infrastructures embedded in the external environment in ways that make possible and require universal uniform (numeric) measurements as essential determinants of the representations of things. The mathematical project of sociocognitive forms of life leads to the development of a certain mathematics in a narrower sense of equations, proofs, axioms, and model-based analytics. But this form of an unmodern science does not arise because mathematics functions as a tool applied to objects in an independently motivated way. Rather, that mathematics, and a particular kind of mathematics taking the form of specific models, could come into play and has to come into play as a consequence of the larger unfolding of the mathematical project as learning through what is already known and what is known in that way.

This sense of the mathematical project's will to ground itself in terms of its own inner requirements as the standard of all thought (Heidegger, 1967, p. 100) was effectively – though implicitly – understood by Bohr when he said that the quantum phenomenon forces us to accept that we are so suspended in language we cannot tell up from down (Petersen, 1968, pp. 187–188; French & Kennedy, 1985, p. 302; Ottaviani & Purvis, 2009). Peirce's (1955, p. 230; 1992; Peirce et al., 1977) assertion that "We have no power of thinking without signs" and Gadamer's comment that "the process of understanding moves entirely in the sphere of a meaning mediated by the linguistic tradition" (1989, p. 391) aptly sum up Bohr's sense of how we are suspended in language. Wheeler then infers what he calls the "deepest lesson" of quantum mechanics, that "no elementary quantum phenomenon is a phenomenon until it is a recorded (observed) phenomenon" (Wheeler & Zurek, 1983, p. xvi; Wheeler, 1980, p. 153; 1981; 1982, p. 201; 1988, 1994, 2014, 2018).

This same theme infuses Scott's (1998, p. 357) call to take language as a model; Ricoeur's (1981) model of the text and paradigm of reading; Derrida's (1976) grammatology; Galison's (1997, 1999) trading zone; Ihde's (1991) readable technologies; Haraway's (1996) feminist diffractions; Heelan's (1983) hermeneutic of instrumentation; Lenoir's (1998) inscription devices; Nersessian's (2002, 2008) account of Maxwell's modeling practices; Star's (1988; Bowker et al., 2015) boundary objects; Wittgenstein's (1958) language games and (Wittgenstein, 1922, p. 74) recognition that "the limits of my language mean the limits of my world"; Gadamer's (1989) sense of language as medium; Dewey's (1954) focus on language as the vehicle of thought, and Sebeok's (2001) and others' (Brier, 2013, 2021; Danesi, 2017; Maran, 2007; Merrell, 1996; Nöth, 2018, 2021) semiotics, ecosemiotics, biosemiotics, cybersemiotics, etc. These latter provide an opportunity for effectively connecting the evolutionary semiotics of complex information flows with the deconstruction of modern metaphysics and the emergence of a new world hypothesis and root metaphor.

Bohr's remark and Wheeler's expansions on it independently motivated Fisher (1988, p. 206; 2003, pp. 806–807) and Sebeok (1991, p. 143; Merrell, 2011, p. 254) to take language as a model. Several related and independent efforts in quantum information theory and semiotics have emerged over the years (Dosch et al., 2006; Galofaro et al., 2018; Jaeger, 2023; Meijer, 2015).

This explication of language as a key factor previously hidden in the background leads to enhanced appreciation for the ways ideas, words, and things are integrated in discourse. Contrary to the long history of various forms of idealism, instrumentalism, operationalism, empiricism, etc., our here taking everyday language as a whole, semiotically, demands that no single one of those levels can serve as the basis for a reduction scheme. Instead, they separate into somewhat interrelated and somewhat independent domains. The complexities of models understood as heuristic fictions, metaphors, boundary objects, and trading zones then follow from taking language as a model for science's integrations of theories, instruments, and data (Ackermann, 1985; Butterfield, 1957; Black, 1961; Hesse, 1970; Galison, 1997; Kuhn, 1993; Bowker et al., 2015). Universally accessible linguistic systems and metasystems integrating these concrete, abstract, and formal levels return us to the model of hierarchical complexity with which we started, and which now sets up a semiotic perspective on a shared concept system for measurement and metrology spanning the arts and sciences (Fisher, 2003, 2004, 2020, 2021a, 2023a/b).

Consciousness understands consciousness, then, when the role of language as the vehicle of thought ceases to be ignored as inconsequential and is instead identified as a model necessary and sufficient to purposes of caring for sociotechnical forms of life as we do our children (Latour, 2012), from the moment of conception through gestation to midwifery, nurturing to maturity, and eventual passing (Fisher, 2007, 2009a, 2010a, 2011a). The pragmatic operationalization of consciousness understood for what it is will be achieved by means of models informing systematic, metasystematic, paradigmatic, and cross-paradigmatic reproductions of predicted effects (De Boeck & Wilson, 2004; Embretson, 2010; Fischer, 1973; Green & Kluever, 1992; Smith, 1996b; Stenner et al., 2013) throughout socially distributed, quality assured metrological networks (Fisher & Stenner, 2016; Mari & Wilson, 2014; Mari et al., 2023; Pendrill, 2014, 2019; Pendrill & Fisher, 2015) informing multilevel communications, legal, and economic processes (Fisher, 2012c, 2023a). Though these semiotic metalanguages and metacommunications may be cast as forms of artificial intelligence, they are paradigmatically distinct in ways that make them more aptly termed forms of natural intelligence (Barney & Barney, 2024; Barney & Fisher, 2017).

Prigogine and Allen's (1982) concern with understanding evolving complexity applicable across physics and psychology implicitly takes up Wheeler's mystery of when consciousness will understand consciousness. Prigogine and Allen (1982, pp. 24–28) introduce three approaches to change in physico-chemical systems generalizable across all evolving forms of life. These are:

(a) a phenomenological-empirical approach based on the Poisson distribution, as is also recognized by Rasch (1960, chapter 2; 1977, pp. 62–65, 92) as a necessary condition for specifically objective comparisons (also see Meredith 1968, Falmagne & Doignon 1997, and Graßhoff et al., 2020);

(b) a stochastic approach, using an example of a Markov process in an individual-level probabilistic model analogous to Rasch's, as is further developed by Bartolucci et al. (2011), Falmagne and Doignon (1997), Graßhoff et al., (2020), Karabatsos & Batchelder (2003), Meredith (1968), and so on; and

(c) dynamic laws conceived not in terms of stochastic descriptions assuming a resigned acceptance of probabilities as making do with incomplete information denoting only ignorance and imprecision, as when it is necessary to sample from large populations, but in terms that "may express some basic characteristic of the deterministic laws of nature" (Prigogine & Allen, 1982, p. 5), here in parallel with Rasch's (1961, 1977) models of stochastic invariance, use of Ronald Fisher's (1922, 1934b) distinction between population-level and individual-level statistical sufficiency, and his (Rasch, 1960, pp. 110–115) taking Maxwell's analysis of Newton's second law as the basis of the form for his models.

Prigogine and Allen's interest in developing an alternative sense of probability could well have made good use of Ronald Fisher's (1934b, p. 287) contrast of deductive and inductive perspectives on probability, as Fisher points out that:

> probability is appropriate to a class of cases in which uncertain inferences are possible from the general to the particular, while likelihood is appropriate to the class of cases arising in the problem of estimation, where we can draw inferences, subject to a different kind of uncertainty, from the particular to the general.

What Prigogine is driving at is exactly the same thing as Rogosa's (1987, p. 193) "critical distinction . . . between models that start with the individual process as opposed to models for relations among variables." Duncan and Stenbeck (1988, pp. 24–25) similarly claim that:

> The main point to emphasize here is that the postulate of probabilistic response must be clearly distinguished in both concept and research design from the stochastic variation of data that arises from random sampling of a heterogeneous population. The distinction is completely blurred in our conventional statistical training and practice of data analysis, wherein the stochastic aspects of the statistical model are most easily justified by the idea of sampling from a population distribution. We seldom stop to wonder if sampling is the only reason for making the model stochastic. The perverse consequence of doing good statistics is, therefore, to suppress curiosity about the actual processes that generate the data.

Falmagne and Doignon (1997, p. 135) focus on the response processes generating data when they present a

> theory purporting to explain how rationality could evolve from a naive state portrayed by the empty relation, to a sophisticated state represented by a semiorder or some other kind of order relation. This evolution was formalized by a stochastic process with three interlinked parts. One is a Poisson process governing the times t1; t2; . . .; tn; . . . of occurrence of quantum events of information, called tokens, which are delivered by the medium. The second is a probability distribution on the collection of all possible tokens, which regulates the nature of the quantum event occurring at time n. Any token is formalized by some pair xy of distinct alternatives, bearing a positive or negative tag. The occurrence of a positive token xy signals a quantum superiority of x over y, while the corresponding negative token y indicates the absence of such a superiority. The last part of the stochastic process is a Markov process describing the changes of states occurring in the subject as a result of the occurrence of particular tokens.

The ratio of rationality emerges via the evolutionary process from an empty relation to an order relation in a manner that maps into the metaphoric process, as shown in empirical survey studies of the "love is a rose" and "life is a mango" metaphors (Fisher, 1988, 1990, 1995, 2011b, 2012a). The invariance of the agreeability and disagreeability of the metaphors' entailments establishes the semiotic basis for scientific models as heuristic fictions, as unrealistic idealizations connected to unique local circumstances by standardized word forms. Language mediates learning at the individual level in the way pre-existing, validated words spelled in a familiar alphabet and pronounced in already-grasped phonemes are used in developing understandings of known and stable conceptual-empirical associations. Collective learning also proceeds by applying what is already known to something new and not yet known, but now the level of complexity has shifted to a re-ordering of a familiar word's commonplace associations, with the aim of identifying a new model sufficient to the tasks of reason and communication.

Sufficient reason is then an outcome of languages' self-organized semiotic modeling processes, processes that are implicitly statistically sufficient, and reason becomes sufficient at new higher order levels of complexity when semiotic models are explicitly formulated to be statistically sufficient. Much the same kind of result has been obtained in developmental psychology (Commons et al., 2014; Dawson-Tunik et al., 2005; Fischer & Farrar, 1987; Overton, 1998, 2015). Language serves as the vehicle of thought to the extent it extends physically embodied understanding and enables intuitive flows of associations, in a manner akin to viral contagions (Commons & Goodheart 2008, p. 413; Cozzo et al., 2013; Pastor-Satorras et al., 2015; Platt, 1961) not just metaphorically, but physically and neurologically (Danilov & Mihailova, 2021; Hodas & Lerman, 2014; Theriault et al., 2021).

Overton's (2015, pp. 31–33) development of Leibniz's theme of the identity of opposites in a synthetic coordination of physical and social systems then validates Wheeler's (1974, p. 689) suggestion of a "Leibniz logic loop" that leads from physics (Bohr's emphasis on complementarity between question and answer as requiring being suspended in language) through participant observation back to human consciousness and participatory partnership with natural phenomena in a living universe.

6.4.2 The mystery of the universe: "The universe runs its course from big bang to collapse; but what part do the future requirements for life and mind have in fixing the physics that comes into being at that big bang?"

Wheeler's (1974) second question provokes a question in response: Do the future requirements for life and mind fix the physics that come into being at the big bang? Or is this a teleological way of putting the cart before the horse? Unless one desires to posit the untestable hypothesis of an omniscient creator designing physical properties to satisfy the requirements of life and mind, does it not seem essential for the physics to fix the future requirements for life and mind, instead of vice versa? As Smolin put it, "The Cosmological questions such as Why these laws? and Why the initial conditions? cannot be answered by a method that takes the laws and initial conditions as input" (Smolin, 2014, p. 250).

Taking the physics as the point of departure, answers to Wheeler's questions can be discerned in the works of physicists such as Bohm (1980), Bohm et al. (1987), and Prigogine (1997). The recurring theme concerns the existence of quantum properties involving a play of interactions that have the same complementary form as interactions involving human participant observers. As Bohm et al. (1987, p. 327; also see pp. 334, 337, 340) says:

> All sorts of quantum processes, such as transitions between states, fusion of two systems into one and fission of one system into two, are able to take place without the need for a human observer.
>
> . . . an objective quantum ontology is possible, in which the existence of the universe can be discussed, without the need for observers or for collapse of the wave function.

Prigogine focuses on quantum level stochastic resonances (Large Poincare Systems; LPS) and irreversible processes, arriving at much the same conclusion as Bohm, saying that the LPS "measure themselves" in a way that is indistinguishable from a human observer's measurements. This means that "the observer no longer plays some extravagant role in the evolution of nature" and that

> In this sense, our approach restores sanity. It eliminates the anthropocentric features implicit in the traditional formulation of quantum theory. (Prigogine, 1997, p. 151)

Contrary to mystical invocations of human observation as somehow magically creating the universe, in the manner of Wigner (1960) or von Neumann (1955), Bohm and Prigogine, like Wheeler, see the universe as both objective and participatory (also see Khrennikov, 2020; Khrennikov & Basieva, 2023). The elimination of human subjectivity as an anthropomorphizing force creating reality is also taken up in semiotics (Brier, 2010, p. 1907; Brier, 2011, p. 46; Herrmann-Pillath, 2010).

Pattee (1979, 2012) adopts Bohr's sense of complementarity in characterizing biological processes as interacting mutual measurements and cites Prigogine in this regard. Prigogine and Lefevre (1973, p. 124) also seek to advance biology by aiming to "provide a physico-chemical basis of evolution towards structures of increased complexity."

These ideas frame a context in which Wheeler's question about the requirements of life and mind can be seen in the physics that came into being at the big bang. Bohm accordingly theorizes about a new science based in a holistic, non-Cartesian "world view in which consciousness and reality would not be fragmented from each other . . . . [so that that] world view is itself an overall movement of thought" performing a kind of "dance" (Bohm, 1980, p. xii). Bohm here implicitly echoes Hegel, who traced out "the self-movement of the concept" in the unity of thing and thought (Gadamer, 1976, p. 11). This theme also recalls the ancient Greek sense of method as the movement of things experienced in thought (Gadamer, 1989, pp. 104, 460, 474; Heidegger, 1991, p. 63; Gasché, 2014). Hegel's logic thinks

> of change and transformation in their dynamic flux not by fixating movement in abstract static descriptions but by *performing movement itself*. By bringing change to bear directly on pure thinking, by making thinking one with the movement it accounts for, Hegel's logic *does* the very thing that it purports to understand. (Nuzzo, 2018, p. 5)

Nuzzo (2018, p. 4) then contends that "Hegel's logic is the crucial intellectual tool that can help us weave the elusive stories of our own present . . .." Importantly, Gadamer (1989, p. 104) stresses the qualities of play infusing the experience of unified things and thought in the structure of dialogue.

Potentials for coalescing a nondualistic paradigm for the sciences are then further suggested by Prigogine when he describes a "new dialogue with nature" characterized not by centralized control and management but by participation in a self-organized and evolving flow (Prigogine, 1997, pp. 71, 154–162; Prigogine & Stengers, 1984, p. 22). Gadamer (1989, p. 367) implicitly concurs:

> To conduct a conversation means to allow oneself to be conducted by the subject matter to which the partners in the dialogue are oriented. It requires that one does not try to argue the

> other person down but that one really considers the weight of the other's opinion. Hence it is an art of testing. But the art of testing is the art of questioning. For we have seen that to question means to lay open, to place in the open. As against the fixity of opinions, questioning makes the object and all its possibilities fluid. A person skilled in the "art" of questioning is a person who can prevent questions from being suppressed by the dominant opinion. A person who possesses this art will himself search for everything in favor of an opinion. Dialectic consists not in trying to discover the weakness of what is said, but in bringing out its real strength. It is not the art of arguing (which can make a strong case out of a weak one) but the art of thinking (which can strengthen objections by referring to the subject matter). The unique and continuing relevance of the Platonic dialogues is due to this art of strengthening.

Science is a fulfillment of the art of strengthening understanding by constantly deferring to the object of investigation, allowing it to assert its own independent voice in the conversation. Listening of this kind often requires a sensitivity to unexpected results and an ability to recognize the value of an answer to a question that was not actually asked, a demand that can make persisting in the art of questioning inordinately difficult. The key importance of being able to pivot off of preconceived ideas toward new imaginary possibilities is evident in the repeated instances in the history of science in which the real strengths of various phenomena were revealed, as in the discoveries of penicillin, vulcanized rubber, post-it note glue, X-rays, and smart dust.

Past the positive productivity of applied science introducing new technical effects into everyday life via standardization, Cook (1914/1979, pp. 426–436), Kuhn (1977, p. 205), and Rasch (1960, p. 124; 1972/2010, p. 1254) all saw the primary value of measurement standards in research to follow from the capacity to reveal anomalies. Cook (1914/1979, pp. 428, 430) held that natural laws "are the instrument of science, not its aim," and that "the whole value . . . of any law is that it enables us to discover exceptions." Kuhn (1977, p. 219) similarly wrote that "To discover quantitative regularity one must normally know what regularity one is seeking and one's instruments must be designed accordingly," and Rasch (1960, p. 124) wrote that "Once a law has been established within a certain field then the law itself may serve as a tool for deciding whether or not added stimuli and/or objects belong to the original group." As Kuhn said, precision measurements reveal the departure from expectations with remarkable finesse and may serve to provoke further investigations into them. Rasch (1960, p. 10) noted that his models for psychological and social measurement would be useful for illuminating exceptions to the rule in the same way that consistent inconsistencies in the orbit of Uranus led to the discovery of Neptune, the same example given by Kuhn (1977, p. 205) and by Cook (1914/1979, p. 431).

Latour (2004, p. 217), citing work co-authored by Stengers, suggests that the social sciences may become as scientific as the natural sciences to the extent they devise their inquiries to maximize the recalcitrance of the phenomena investigated. As Gadamer (1981, p. 164) puts it, "the fruitfulness of scientific questioning is defined in an adequate manner if it is really open to answers in the sense that experience can refuse the anticipated confirmation." A capacity for learning and growth hinges on being able to yield to certain kinds of experiences, such that a more productive outcome may follow from avoiding, rather than engaging in, forceful assertions. “Success is then not a question of how unchanged the self emerges from the test nor how much it has bent the nonself to its will, but how enriched it became in the process” (Bettelheim, 1967, p. 81).

The capacity to export technical effects from the laboratory into the outer world depends, after all, on understanding the practical limits on the conditions in which those effects will reliably persist (Ihde, 1991, pp. 133–135). In addition, adapting to local circumstances without compromising global comparability requires ways of keeping questions open to the demands of varying contexts; cutting off questioning prevents those in dialogue from being able to arrive at shared understandings and may impose extraneous concerns. Testing hypotheses as to the nature of what is measured then implies just the kind of deference to the manifestation of the construct that comprises tests of the strength of the object of the dialogue or openness to recalcitrant responses from it.

Here we encounter the crucial role in measurement modeling played by evaluations of data consistency (Smith, 1991, 1996a) and the predictive power of explanatory models (De Boeck & Wilson, 2004; Embretson, 2010; G. Fischer, 1973; Green & Kleuver, 1992; Smith, 1996b; Stenner et al., 2013). In contrast to statistical methods prioritizing the descriptive power of models in relation to data, where the model selected may in fact overfit the data and have no generalizability (Bamber & van Santen, 1985; Duncan & Stenbeck, 1988; San Martin et al., 2015), scientific measurement evaluates the model-data fit in relation to criteria of meaningfulness, on the basis of the GIGO (Garbage In, Garbage Out) principle (San Martin et al., 2024; San Martin & Rolin, 2013).

This reversal of the usual statistical procedure in measurement practice has been controversial (Andrich, 1989, 2004; Embretson, 1996; Wright, 1984) because of the perception among some (Hambleton et al., 1991; Weiss, 2021) that inordinate value is being placed on a model possessing what are considered to be convenient but expendable properties involving simplistic, stringent, and unrealistic assumptions. Gadamer captures the philosophical crux of the difference in perspective characterizing this controversy, saying:

> In contrast with the modern procedure of verifying a hypothesis, the hypothesis of the *eidos* [an abstract ideal] is not to be tested against an “experience” which would validate or invalidate it. Such a procedure would be totally absurd in respect to a postulated *eidos*: that which constitutes being a horse could never be proved or disproved by a particular horse. Instead, the test which is to be applied in respect to the *eidos* is a test of the immanent, internal coherence of all that is intrinsic to it. One should go no further until one is clear about what the assumption of the *eidos* means and what it does not mean. It should be noted that consequently the hypothesis is not to be tested against presumed empirical consequences, but conversely the empirical consequences are to be tested against the hypothesis, i.e., that from the start everything empirical or accidental which the *eidos* does not mean and imply is to be excluded from consideration. This means above all that the particular which participates in an *eidos* is of importance in an argument only in regard to that in which it may be said to participate, i.e., only in regard to its eidetic content. All logical confusion is a consequence of failing to distinguish and separate the *eidos* from what merely participates in it.

Measurement model fit analyses proceed in a manner that is remarkably well described by Gadamer. This apt characterization is possible—even though Gadamer likely never estimated the parameters of a measurement model in his life—because of the way he speaks to the concerns of meaningful identity. Measurement modeling involves the prescription of the kinds of data structures necessary and sufficient to the estimation of a unit quantity that remains invariant across samples and instruments. The process of testing the strength of the object of inquiry requires exactly the kind of concern for unity and sameness Gadamer describes. Where descriptive statistical modeling takes responses to the questions asked from the persons responding as inherently objective in and of themselves, prescriptive measurement modeling instead looks to the qualitative participation of the questions and answers in the unfolding of the object of discourse. Here, the Hegelian activity of the thing itself as experienced in thought becomes the fundamental basis for a methodically replicable process (Fisher, 2004), instead of a set of rules presumed to lead to a predetermined end in the modern sense of statistical hypothesis testing.

At the same time, it must be emphasized that the "modest witness" who defers to the object of inquiry and erases their participation in the production of scientific effects should not go so far as to perpetuate a metaphysics elevating a pure and disinterested epistemological agent at the expense of others rendered both invisible and voiceless (Haraway, 1996). Instead, to more fully constitute a new paradigmatically distinct form of "strong objectivity" in which "embedded relationality is the prophylaxis for both relativism and transcendence" we must more fully maximize the recalcitrance of the phenomena in the ways we deal with exceptions to the rule that resist standardization (Haraway, 1996, pp. 438–440). Haraway draws on Harding (1991, 2008) and Star (1988; Bowker et al., 2015) in this vein in ways that complement Galison's (1997, pp. 843–844) search for a stochastically resonant metaphor of how it is "the disorder of the scientific community – the laminated, finite, partially independent strata supporting one another" and the "*dis*unification of science – the intercalation of *different* patterns of argument – that is responsible for its strength and coherence." The diffraction patterns emphasized by Haraway seem to be excellent candidates exemplifying both the noise-induced order described by Galison and the stochastic invariances structuring probabilistic measurement (Fisher, 1992b, 2011c; Fisher & Wilson, 2015).

In this complexity we see how physics as a discipline and as nature fixes the requirements for life and mind by being playful. Nersessian (1998) draws out the playful dynamics of experimental science. Participants in dialogue give themselves over to the play of language games in a manner that echoes the play of nature seen in light through the leaves, waves on the beach, recombinant DNA, or the pretend fighting of puppies. Gadamer (1989, pp. 104, 107) then points out that it is more appropriate to say that humanity plays naturally than it is to say that nature plays like humans. As he (1989, p. 108) says, playful "self-presentation is a universal ontological characteristic of nature." Here we see the crux of how the physics of nature's playful dance fix the requirements for life and mind, where the movement of thought dances not only in people's minds but also in flows of matter, energy, and information throughout the universe.

Wheeler's notion that the universe could be a home for human life and mind finds further support when Prigogine and Stengers say they conceive of knowledge as both objective and participatory, and "believe that we are heading toward a new synthesis, a new naturalism," one that

> does not suppose any fundamental mode of description; each level of description is implied by another and implies the other. We need a multiplicity of levels that are all connected, none of which may have a claim to preeminence. (Prigogine & Stengers (1984, p. 300)

Formal theory and axioms, abstract measurement standards, and concrete data are similarly irreducible levels of description that imply one another in ways that do not allow any one of them to serve as the ground of a homogenizing logic. Feynman (1965, p. 46) and Toulmin (1961, pp. 28–33) both spoke in this vein of the need in science for both Babylonian algorithmic and Greek axiomatic forms of thinking (Niederée, 1994, p. 583). Star's (Bowker et al., 2015) investigations of standards implementations led to the concept of the boundary object, which is simultaneously abstract and concrete in the way it both facilitates global communications and the negotiation of locally situated meanings unique to specific circumstances.

In the wake of his ethnographic studies of microphysics' communities of theoreticians, instrument makers, and experimentalists, Galison (1999, p. 143) similarly proposed an "open-ended model" he calls a "trading zone" that:

1) is "tripartite in allowing partial autonomy to instrumentation, experimentation, and theory;"
2) allows each area its own break points in the evolution of its ideas and methods;
3) asserts that abrupt changes are not likely to occur across areas at the same time; and
4) leads us to "expect a rough parity among the strata – no one level is privileged, no one subculture has the special position of narrating the right development of the field or serving as the reduction basis."

Finally, the model of hierarchical complexity (Commons & Richards, 2002; Dawson-Tunik et al., 2005; Fischer & Farrar, 1987) mentioned in the context of Wheeler's first question systematically theorizes the developmental and evolutionary processes involved in moving through and integrating stage transitions over the lifespan and through the history of science (Commons & Bresette, 2006; Commons et al., 2011).

6.4.3 The mystery of the quantum: "The quantum principle says, 'No physics without an observer'; but from what comes the necessity of this principle in the construction of the world?"

The answer to Wheeler's third question has already emerged in the responses to the first two. The necessity of the quantum principle (no physics without an observer, which need not be human) in the construction of the world follows from the condition of physical existence: that the physics of matter, energy, and information emerge and are real only to the extent that all things and processes observe and measure themselves and each other via their interactions. Things define what they are through their interactions. Humanity is simultaneously both a product and an initiator of these interactions in ways that make it the eyes of the world. But we would do well to join Wheeler (1974, p. 689) in replacing the primacy of the observer with a more contextualized participator.

The necessity of "no physics without an observer" comes about because the nature of the quantum phenomenon exceeds the epistemological limits of modern Western dualist metaphysics. The microscopic scale of the quantum phenomenon requires a technical macroscopic apparatus to make it visible, perceptible, and legible, necessitating the overthrow of the modern reification of direct perception as constituting meaning. The always already existing unity of things and thought in language long ignored by the Western worldview can no longer be brushed aside in the quantum context. Because the modern worldview perpetuates the Pythagorean confusion of numeric and geometric figures with existence (Gadamer, 1980, p. 35), the dependence of the quantum phenomenon's manifestation on the macroscopic form of the question asked caused considerable consternation. But the main consequence of the fact that the elementary quantum phenomenon is not a phenomenon until it is registered on a recording device could possibly be, following Wheeler's speculations, that humanity may find itself more suitably feeling at home in a participatory universe.

That feeling of being at home, of being included as a fundamentally integrated part of the universal whole, will be consummated in an extended SI made coherent by its incorporation of the nonequilibrium thermodynamics of entropy dissipation as a defining factor. Theories of natural, psychological, and social evolution are increasingly combined within a semiotic framework structuring hierarchically complex flows of matter, energy, and information (Brier, 2021; de Castro & McShea, 2022). Accordingly, in the same way that the speed of light is implicated across multiple units in the existing SI, so, also, will entropy be implicated in the learning progressions, developmental sequences, and healing trajectories of an extended SI's social and psychological units (Fisher, 2024). Here, the play *is* the thing. Humanity, too, plays, absorbed into the back-and-forth flow of existence in exactly the same ways as gamboling lambs or the music of bird song.

A long history of nondualist alternatives to subjective experience alienated from an objective reality offer new possibilities for creative exploration. The language of modern science ignores the semiotic levels of complexity in ways that entail constantly confusing them for each other, and inappropriately dragging them from one to another in variations on the ecological and atomistic fallacies. Statistical data reports are summarily removed from the concrete level and – in a blatant example of the ecological fallacy (Alker, 1969; Gnialdi et al., 2018), referred to by Whitehead (1925, pp. 52–58) as the fallacy of misplaced concreteness – are considered abstract quantities even when their ordinal status is undisputed, no unit value is defined or named, and no explanatory model predicts values reproducible from theory. Nondualistic, non-Western, nonmodern, or premodern metaphysics accord with the seemingly exotic characteristics of quantum mechanics in what many may find to be surprisingly accessible ways. The transformative shift taking place is one in which the new and original is simultaneously recognized as old and familiar in a satisfying re-enactment of the saying, "the more things change, the more they stay the same." Instead of intimidating complications or insurmountable obstacles, what we find are intensifications of intimately familiar ideas and experiences.

6.5 Aesthetics: beauty and meaning

6.5.1 Motivations for doing science

Many scientists do not think first of factual truth or practical utility when asked what they value in science. Instead, they extol beauty as their primary motivation for doing science, with truth and usefulness contingent on this more compelling source of interest in nature. Chandrasekar (1979, p. 27) and Townes (2001, p. 299)—first author on Ben Wright's first publication (Townes, et al., 1948)—two prominent, Nobel-winning physicists, approvingly quote Keats' "Ode on a Grecian Urn": "Beauty is truth, truth is beauty." Both Chandrasekar (p. 25) and Townes (p. 298) also point toward Poincaré's emphasis on the aesthetic attractiveness of simplicity and immensity in nature. Poincaré (quoted in Chandrasekar, 1979, p. 25) claims that:

> The Scientist does not study nature because it is useful to do so. He studies it because he takes pleasure in it; and he takes pleasure in it because it is beautiful. If nature were not beautiful, it would not be worth knowing and life would not be worth living.

McAllister (1990) offers a vigorous defense of Dirac's (1963, p. 47) claim that "It is more important to have beauty in one's equations than to have them fit experiment." The hermeneutic philosopher, Gadamer (1998, p. 73) similarly held that "science exists and is important for no other reason than because it is beautiful." Keats expands, saying "what the imagination seizes as beauty must be truth," leading Chandrasekar (1979,

p. 27) to write, “what is intelligible is also beautiful.” Gadamer (1989, p. 490) independently concurs, saying, “when we understand a text, what is meaningful in it captivates us just as the beautiful captivates us.”

The creativity involved in bringing the experience of beauty into words led the mathematician Weierstrass to say that “No mathematician can be a complete mathematician unless he is also something of a poet” (quoted in Huntley, 1970, p. 1). Bronowski then similarly held that mathematics “is a literature in its own right . . . a form of poetry, which has the same relation to the prose of practical mathematics as poetry has in relation to any other language” (quoted in Huntley, 1970, p. 3). Another mathematician, Hardy, also valued this connection, saying that “the mathematician’s patterns, like the painter’s or the poet’s, must be beautiful” (quoted in Huntley, 1970, p. 84).

Though this aesthetic dimension does not typically attract much attention in the philosophy of science, a clear account of it is necessary to addressing the inescapable and unresolved problem of metaphor. Logical thinkers since Plato have sought to exclude poets and rhetoric from serious discourse because metaphor says one thing (love is a rose) but means another (love has qualities of beauty, thorniness, color, fragrance, etc. analogous to those of roses). These efforts at rigorous logic ultimately can only assert internally inconsistent positions in which metaphors and rhetoric are employed in arguments against their use. Locke (1979, p. 508), for instance, eloquently argues against eloquence, all the while persuasively relying (Locke, 1979, pp. 509, 510, 578) on the metaphor “language is a conduit” (Cohen, 1979, pp. 2–3). Paradoxically, it proves counterproductive to rigidly adhere to the apparently simple requirements of logically consistent identities (this is this and that is that), noncontradictory assertions (this is this and cannot be that), and the excluded middle (no gray zone of acceptable variation allows this to fade into that) (Estep, 1993; Heelan, 1974; Keller & Tian, 2021; van Fraassen, 1974, 2008).

But metaphors are unavoidably implicated in science and mathematics in even more elementary and foundational ways. Plato (in the *Phaedo*, 96b) noticed that counting abstracts unrealistic similarities from concrete and unique things, situations, processes, and people (Ballard, 1978, pp. 186–190). Plato also redefined the elements of geometry as idealizations (points as indivisible lines, lines as indivisible planes, etc.) for the express purpose of resolving the Pythagorean crisis induced by the irrationality of pi in estimating the circumference of a circle, and of the square root of 2 in the hypotenuse of the right isosceles triangle (Boyer, 1949, p. 18; Cajori, 1999, pp. 25–26; Gadamer, 1980, pp. 35, 100–101). Plato ironically saw how to resolve the Pythagorean crisis of irrationality by idealizing all mathematical relationships of number and geometry and distinguishing between figure and meaning, name and concept, but then did not recognize the metaphors involved in counting unlike units and drawing lines that are actually divisible, not indivisible, planes.

Physics similarly asserts metaphorical fictions, as when Newton’s first law holds that a body left entirely to itself moves uniformly in a straight line, or when Galileo posited perfectly spherical balls rolling on a frictionless plane. No bodies in the physical universe are ever not acted on by external forces, and straight lines ultimately have no basis for existing in a context of curved space-time.

The pragmatic utility obtained from these metaphors is, however, not only undeniable but also ought to compel energetic and rigorous explorations of ways in which that utility can be explained and expanded into applications in new domains. Philosophers have long puzzled over the paradox of sciences supposedly based in experiences of objective facts but which simultaneously assert fundamentally unrealistic representations as laws embodied by those facts (Black, 1962; Bundgaard, 2019; Brown, 2003, p. 195; Butterfield, 1957, pp. 16–17; Cartwright, 1983; Heidegger, 1967, pp. 78, 89–91; Hesse, 1970; Holton, 1988, pp. 42–43). Ways forward resolving this paradox have been difficult to discern, however.

Beauty suggests viable new directions to explore. Beauty is not only cited by many mathematicians and scientists as a primary motivation, it provides unique insights into understanding how to think about meaningful communication in everyday life as well as in science. These insights can be traced from the lessons beauty teaches concerning meaning in language, following from Diotima's story of Eros recounted by Socrates in Plato's *Symposium*. Diotima sets up a general theory and practice of meaningful communication encompassing everyday and scientific discourse in an overarching semiotic model. When unrealistic heuristic fictions are expressed via linguistic standards and are applied in negotiating unique local circumstances, we obtain an erotic idealism that takes language as the vehicle of thought.

Whitehead (1925, p. 107) remarks on the way that the quantum revolution in physics inspired new imaginative possibilities in the minds of scientists not because they individually acquired new capacities for visualization but because of the availability of new instrumentation. Language is again implicated as the medium in which we think when Whitehead elsewhere notes that civilization does not progress as a result of clear thinking on the part of individuals, but because of the distribution of capacities for executing operations we do not understand (Whitehead, 1911, p. 61). In the same way that we all make use of everyday languages' signs, symbols, semantics, syntaxes, and grammars with little or no understanding of the seemingly arbitrary reasons why or how they are structured as they are, so, too, do we routinely make use of clocks, thermometers, smartphones, and machines none among us is capable of explaining, inventing, or manufacturing on our own.

This is the domain of problems involved when scientists say that scientific language is nothing but an extension and refinement of everyday language (Einstein, 1954, p. 290; Bohr, 1963, p. 9). Semiotic models of the unity of thing and thought (Brier, 2013, 2021; Danesi, 2017; Maran, 2007; Nöth, 2018; Olteanu, 2021; Sebeok, 2001) set up new opportunities for advancing science and do so in ways that do not repeat the error of modeling psychology and the social sciences operationally on the natural sciences, but which instead tap into the roots of scientific thinking in everyday language and feed technical complexities forward into everyday language usage (Fisher, 2020, 2023a).

The pragmatism involved in accepting the divergence of the concrete from the formal and abstract is captured in Coleridge's characterization of beauty as a function of unity in variety (quoted in Huntley, 1970, pp. 14, 85). Desire for the beauty of the beloved violates the law of the excluded middle in the same way that language suspends us between the perfect form of idealized meanings and the concrete realities of local circumstances (Gelven, 1984). Pragmatically integrating formal ideals, abstract media, and concrete things in semiotic systems, metasystems, and paradigms (Commons & Bresette, 2006; Commons et al., 2011) makes scientific thinking and language widely available throughout socially distributed networks, enabling us to refine and make explicit the mathematical structures tacitly embedded in everyday thinking and language (Fisher, 2020, 2021, 2022a/b).

Bringing the implications of beauty for meaning to bear in this way sets up new capacities for donning what Butterfield (1957, p. 17) called "a different kind of thinking-cap, a transposition in the mind." This shift in perspective involves imagining a geometry of relationships comprehensible in the terms of answers to questions of one's own devising. As Kant put it and as Heidegger (1967, p. 68) elaborated, the amount of genuine science found in any domain of investigation is a function of its mathematical sophistication. The typical "thinking-cap" taken for granted in most scientific realism naively persists in the futile expectation that the description of observed facts will lead to the discovery of laws. But the history of science documents no instances in which repeated photographic methods of observation accumulate into laws, or of rules that consistently lead to new discoveries when followed (Holton, 1988; Kuhn, 1970; Polanyi, 1974, p. 323; Russell, 1948, pp. 381–386).

Instead, history documents complex interplays between idealized models and technologically embodied, repeatable observations (Ihde, 1991; Latour, 1983, 1987, 2005; Price, 1986; Shapin & Schaffer, 1985; Wise, 1995). The "different kind of thinking-cap" that involves "a transposition in the mind" took place in astronomy when Copernicus shifted from the Ptolemaic descriptions of planetary epicycles to a geometry of the heavens that included the earth (Burtt, 1954, p. 39). Maxwell effected a similar shift in his electromagnetic studies (Nersessian, 2002, 2008).

Everyday language is extended into science via modeling processes structured similarly enough across the natural and social sciences (Fisher, 2010d; Fisher & Stenner, 2013) to support the emergence of a new cross-disciplinary theory and practice of measurement and metrology (Cano et al., 2019; Fisher, 2020a, 2021; Fisher & Cano, 2023; Fisher & Stenner, 2016; Mari & Wilson, 2014; Mari et al., 2023; Pendrill, 2019; Pendrill & Fisher, 2015).

In this context, the pragmatic value of foregrounding the experience of beauty then can be extended into systems of theoretical predictions embodied in standardized technical media structuring coordinated locally situated applications. The story of Eros told to Socrates by Diotima points the way.

6.5.2 Diotima, Eros, and the law of the excluded middle

In Plato's *Symposium*, Socrates recounts his conversation with Diotima, who tells him that Eros, conceived at a feast celebrating the birthday of Aphrodite, is the child of the god of wealth, Poros, and of a mortal woman symbolizing poverty, Penia. As the child of these opposites, in love the beauty of the beloved is desired in a way that violates the law of the excluded middle: no amount of possession removes the desire, and no distance apart removes the feeling of closeness. In love, desire for the beauty of the beloved is never fully satisfied or lost.

Diotima's description of the experience of desire for and love of beauty teaches us how meaning can be thought about as a kind of pragmatic or erotic idealism:

> For, Plato argues, in love we are both ignorant and wise, finite and infinite, possessing and lacking. The lover, in longing for his beloved, cannot be said to possess nor to lack what he desires, since he would not love if he totally lacked, nor would he be able to desire if he totally possessed. (Gelven, 1984, p. 132; Irigaray & Kuykendall, 1988; Nye, 1989, 2015; Orr, 2006)

Love then embodies simultaneously aspects of both possession of the beloved, and the longing for a consummation that can never be realized.

Meaning similarly demands acceptance of the apparent paradox of an asserted ideal never brought into view as a concrete presence. "We neither possess the perfect form of meaning . . . nor are we unaware of it" (Gelven, 1984, p. 132). In the *Symposium*, at least, Plato characterized understanding as judging inevitably flawed and finite perceptions relative to ideals of infinite perfection. Accepting the roles of beauty and words as abstract media integrating unrealistic ideals with concrete actuality in thought is, Gadamer (1989, p. 481) holds, the metaphysical crux of Plato's mathematical philosophy, constituting a kind of pragmatic or erotic idealism. Gadamer (1989, p. 490) expands on the point, saying,

> When we understand a text, what is meaningful in it captivates us just as the beautiful captivates us. It has asserted itself and captivated us before we can come to ourselves and be in a position to test the claim to meaning that it makes. What we encounter in the experience of the beautiful and in understanding the meaning of tradition really has something of the truth of play about it. In understanding we are drawn into an event of truth and arrive, as it were, too late, if we want to know what we are supposed to believe.

Captivation with beauty and meaning infuses language use in ways that make even the simplest and seemingly most certain mathematical and geometrical truths inherently uncertain and subject to doubt. Routine communications allow habitual, automatic associations embedded in language use to efficiently connect intentions with circumstances on the fly, in the moment. As will be seen, the downside of this focusing of attention is that other potentially meaningful associations are ignored.

Beauty is cited as a primary motivating value in the sciences as much as the arts (Caraman & Caraman, 2021; Chandrasekhar, 1979; Cole, 1998; Huntley, 1970; MacArthur, 2021). But captivation with beauty has a fluid dynamism about it that contrasts markedly with the way it is sometimes portrayed as a matter of fixed symmetries and proportions. In that vein, and contrary to the longstanding tradition casting Plato as a philosopher of static ideal forms, close reading of his works leads to a very different and far more complex philosophical legacy (Gadamer, 1980, 1989).

This point finds specific relevance in the relation to Diotima's account of Eros in the *Symposium*, where the typical interpretation of her ladder of increasing appreciation for beauty casts her as an idealist in Plato's reductionist mold (Irigaray & Kuykendall, 1988; Schott, 1988). A richer perspective, however, retains the ambiguous mix of wealth and poverty, possession and loss, infinite and finite, formal ideals and concrete reality at each step up the ladder from beautiful bodies to beautiful souls and minds to beautiful institutions and from there to the sciences (Gadamer, 1989, p. 478; Nye, 1988, 2015; Orr, 2006).

Pragmatic and erotic idealism of this kind is implicated even in a science as seemingly concrete as geometry, where the conceptual ideals associated with drawn figures are never actually present in the images. As is well known, for instance, it is impossible to draw a line of a length corresponding to an irrational number like the square root of two, which is the length of the hypotenuse of a right isosceles triangle where the other two sides both have a length of one. The same thing happens when the radius of a circle has a length of one: squaring that radius gives a circumference equaling pi to the power of one, meaning that the distance around the circle is 3.14159 . . ., an irrational number that can be neither measured nor drawn.

This problem contradicted the Pythagorean view of the world as number. Plato resolved Pythagoreanism's crisis of irrationality by redefining the elements of geometry as idealizations (points are indivisible lines, lines are indivisible planes, etc.) (Cajori, 1999, p. 26; Gadamer, 1980, pp. 35, 100–101; Ricoeur, 1965, p. 202), making rational and irrational numbers conceptually equivalent. Plato then required experience in geometry for entry into his Academy because understanding that numeric and geometric figures are not the mathematical relationships they represent is fundamental to philosophy (Gadamer, 1980, p. 101; Heidegger, 1967, pp. 75–76; Kisiel, 2002; Harries, 2010).

We are so habituated to the usefulness of geometric relations and natural laws that it is easy to forget they are not as true as they are useful. Although it is often said that models are never true but must be applicable (Rasch, 1960, pp. 37–38; 1973/2011; Box, 1979, p. 202), the valuable insight expressed in this statement has unfortunately become a cliché not generally appreciated for its real meaning. The laws of science are fictions (Butterfield, 1957, pp. 16–17, 25–26, 96–98; Cartwright, 1983; Heidegger, 1967, pp. 89–93; Holton, 1988, pp. 41–44) that absorb our attention in the way metaphors do (Black, 1962; Gibbs, 2008; Kuhn, 1993). Recognizing that words and instruments embody captivation with heuristic fictions useful in negotiating local meanings provides a pragmatic way of integrating unrealistic axioms, ideals, and theories with concrete things and data. Here, we begin to act on the recommendation of language as the best model for providing the seemingly paradoxical combination of navigable continuity and locally situated meaning we need in systems intended to improve the human condition (Scott, 1998, p. 357).

6.6 Ethics: golden rule measurement

Every human culture sets a model of behavior that accords with what is sometimes called the Golden Rule, which is the idea that everyone should treat everyone else in the ways they themselves would want to be treated. Examples include:

- One should seek for others the happiness one derives for oneself. (Buddhism)
- All things whatsoever ye would what men should do to you, do ye even so to them. (Christian)
- The true rule in life is to guard and do by the things of others as you do by your own. (Hindu Rig Vedas)
- No one of you is a believer until he loves for his brother what he loves for himself. (Islam)
- What is displeasing to thyself, do not do to others. This is the substance of the law. All else is commentary. (Judaism)
- Treat others as thou wouldst be treated thyself. As thou deemest thyself, so deem others. Then shalt thou become a partner in heaven. (Sikhism)
- Regard your neighbor's gain as your own gain and regard your neighbor's loss as your own loss. (Taoism)
- I am because we are; a universal bond of sharing connects all humanity. (Ubuntu)
- Do as you would be done by. (Zend-Avesta)

These kinds of rules set up proportionate ratios of analogies, such that A is to B as C is to D, or that A is to C as B is to D, etc. Measurement similarly defines proportionate correspondences that locate abilities and difficulties on a common scale relative to an ideal model of how infinite populations of people and challenges interact and relate. Today's standards of measurement, however, vacillate in their definition and implementation of ethical standards. Disproportionate relationships are taken for granted as depending on which groups are preferred or not, or fairness is left up to the judgment of empowered individuals. Perhaps, however, an ethics drawing on principled criteria for beauty and meaning could be systematically embedded in institutional standards?

"Philosophy is entirely defined by the desire for meaning" but raising questions and trying to answer them inherently involves conceptual, gestational, and parturitional labor risking violence (Ricoeur, 1974, pp. 95–96; 2020). Even the poetics of metaphor twists meanings to create new perceptions of things in the world. Any act of speaking or writing poses the risk of imposing a premature conclusion. Can anything be done to systematically justify decisions in ways that both support decision-making and keep conversations alive to new questions and open to new answers?

Measurement is an enactment of analogies embodied in instruments. Instruments are calibrated in experiments designed to realize stable and consistent (invariant) relationships. The hypothesis tested is one in which my measurement is evaluated to see if it is in proportion to one test or survey question (item) as everyone else's measurements are to other items. Alternatively, we might state the hypothesis as evaluating the proportion relating your measure to mine in connection to the proportion between one item and another (a:c::b:d), or relative to a given standard (a:b::c:b).

Speaking scientifically, we model the infinite arrays of all possible questions and answers in ways that allow for the introduction of new instances and the falling out of old ones. Each instance in which the Golden Rule is applied must be vigilantly attentive to each moment in the phenomenological (or ontological) method: the justification of the abstractions reducing data to measured estimates, the theoretical appropriateness of the conceptual constructions, and the particulars of deconstructing empirical figure-meaning dependencies encountered en route to a new reduction (Fisher, 2010b; Fisher & Stenner, 2011).

In that methodological frame of reference, humanity might be able to pose and more effectively answer questions as to how we know when we measure up to our own and others' standards. Maybe it will be possible to know how we can treat each other with more proportionate consideration. Perhaps we can create a shared social reality in which we know when implicit reductions and always premature conclusions are justified in any given application of the Golden Rule. Maybe we can learn how to better rely on our words by determining the capacity of language to bear the weight of meaning in the future as well as it did in the past.

Access to language precedes and informs the choice between discourse and violence, the primary ethical decision we make (Ricoeur, 1974, 2020). Having the option to choose discourse over violence depends on being cared for enough to have learned how to represent oneself to others. Gilligan (1982), Ruddick (1989), and Noddings (1984) similarly focus on care as foundational to relationships, while Habermas (1995) takes up considerateness, and Irigaray (1984), the fecund gifts of life bestowed by lovers on one another. Gadamer (1991, p. 61) asserts that "care for the unity and sameness" of meaning is the "first concern of all dialogical and dialectical inquiry."

Of course, caring for meaning in this way is inextricably interwoven with caring for miscommunications, varied perspectives, and differences of opinion in the context of the hierarchical complexity of discourse. These longstanding and essential matters of concern provoke the question as to what can be done to transform and revitalize our institutions, to put them in accord with an ethic of love, care, and considerateness, while we also critically engage in constructive dialogues around difficult issues.

A lack of care for the unity and sameness of meaning characterizes measurement in psychology and the social sciences. We most pointedly do not typically care for these technologies as we do our children, to adopt Latour's (2012) variation on Kristeva's language. Confusing numeric counts with measured quantities (Bateson, 1978; Wright, 1994; Fisher, 2021) repeats the fundamental error of the Pythagoreans, who did not distinguish concrete geometric and numeric figures from their abstract meaning (Boyer, 1949, p. 17; Gadamer, 1980, p. 35; Fisher, 1992a, 2003). Indeed, until the time of the ancient Greeks, representations of numbers are always found in association with the thing counted, and not as an independent abstraction (Ifrah, 1999; Suppes &

Zinnes, 1963, p. 4). Seeing that 2 + 2 = 4 no matter what is counted is a significant intellectual and cultural advance. Though philosophy begins with Plato's mathematical distinction between name and concept (Gadamer, 1980, p. 100), this essential difference has yet to be incorporated into a wide range of sciences' technical languages and communications (Fisher, 1992a, 2003, 2020, 2021).

And so it is that data and methods are referred to as quantitative even in the absence of any concern whatsoever with establishing the existence of a substantive unit amount meaningfully represented by numbers. Widespread assumptions that numeric expression alone suffices for quantification have, furthermore, supported methodically institutionalized accommodations accepting variably sized ordinal units to the exclusion of readily available interval models and methods.

The reasons for the contradictory break between the values overtly espoused in the conduct of research and those actually in evidence are deeply rooted in the pragmatic functionality of the existing institutions' systems of incentives and rewards. The simplest and most direct path to publications, funding, and promotions for researchers is often one that involves accepting and meeting established methodological expectations, even when these are ill-founded (Bakker, van Dijk, & Wicherts, 2012; Cohen, 1994; Michell, 1986; Salsburg, 1985; Sijtsma, 2016). As Bateson (1972, p. 493) puts it, "we are most of us governed by epistemologies that we know to be wrong." Clarifying the viability of new goals and standards is a key part of the ongoing process of advancing the state of the art in scientific communications and in creating new social contexts supportive of much needed innovations.

Each cultural tradition has its own terms for working through these moments in the methodical ways things come into words and has its own roots that must feed its growth into the global canopy. As Ricoeur (1974, pp. 291–292) puts it

> The task is to exercise a kind of permanent arbitration between technical universalism and the personality constituted on the ethico-political plane. All the struggles of decolonization and liberation are marked by the double necessity of entering into the global technical society and being rooted in the cultural past.

Metrology contributes to the struggles of decolonization and liberation by facilitating entry into the global technical culture and by doing so in ways that are rooted in each distinctive culture's past (Allen & Pak, 2023; Mallinson, 2024; Massof & Bradley, 2023; Sul, 2024). Postmodern deconstructions of modern metaphysics have led to unmodern conceptions of intensely personalized custom-tailored measurements reconciled with technological universals. The creative power realized by science's methodical enactments of the ways things are put into words constitutes a recovery of language's original meaning-giving capacities (Gadamer, 1989, pp. 428–436). Aristotle's distillation of logic created an artificial structure in opposition to the metaphoric origins of concepts that ultimately led to the modern textbook sense of methods that rewrite the history of the field in the name of a utilitarian narrative of efficiency (Kuhn, 1970). But the metaphorical status of the laws of science as heuristic fictions is complemented by the consistent regularity of metaphors' systems of associated commonplaces (Black, 1962) and model-based transformations from qualitative to quantitative mathematical structures (Fisher, 1988, 1995, 2012a).

In a similar vein, though the balance scale is taken as a symbol of justice, much more could be done to take that symbol seriously as an actual model for fair and equitable decisions. In the same way that the Parthenon's lack of parallels and orthogonal angles symbolizes unique citizens joined in common cause (Cook, 1914/1979, pp. 325–332; Fisher & Stenner, 2018b; Pollitt, 1972, pp. 74–78), so might measurements enacted billions of times a day come to also incorporate the same integration of apparently opposed generality and specificity. Ricoeur's focus on the choice in favor of discourse over violence emphasizes avoiding premature conclusions in the hope that a globally shared sense of human identity might be fully realized in a way that does not contradict human singularity (Ricoeur, 1974, pp. 90–91, 166). Might it be possible for a social ethic integrating that paradoxical opposition to spring from probabilistic models of open systems, models allowing us to keep questions alive, provisionally informing judgments that can be revised in light of new data? Much needs to be done here to explore as-yet untried options for educational, health, governance, judicial, social, and other institutions, but perhaps the time is right to commence new considerations of possibilities for transformation.

6.7 Concluding recommendation: convening new discussions on measurement and metrology

In conclusion, brief comments on the need to take a metrological measurement perspective in some key high level areas are in order. Ongoing human, social, and environmental catastrophes urgently demand collective responses coordinating the behaviors and decisions of billions of people. International metrological initiatives could enable the creation of new common markets for human, social, and natural capital. Making it possible for everyone everywhere to think global and act local in new ways could inspire significant new entrepreneurial investments in needed innovations. Instead of continuing to apply the same models and methods in maddening exacerbations of existing problems, isn't it time to break the mold and try approaches that tap proven ideas in original combinations? I briefly address three examples of high-profile measurement applications that could benefit from better informed theory and practice: The Basel Committee on Banking Supervision, the United Nations Agenda 2030 and Sustainable Development Goals, and The Commission on the Measurement of Economic Performance and Social Progress. None of these examples of efforts aimed at improving the human condition pass Scott's (1998, pp. 355–357) integrity test; see Section 6.2.3 for what the test involves, and how measurement systems can be designed to surpass its requirements.

The overall theme involves a contrast between the cross sectional and static ways in which five conditions of sustainable change are conceptualized and put into play when measurements are modeled in terms of numeric counts vs the longitudinal and fluid possibilities opened up when measurements are calibrated as meaningful quantities (Fisher, 2022b). In each example, the vision, plan, resources, skills, and incentives informing the definition of a problem and its possible solutions are unnecessarily compromised by the ways in which measurement methods impact organizational methods. Disconnects between merely numeric methods of quantification and the management of relevant matters of concern in the three existing systems described result in unintended consequences negatively impacting potentials for achieving the desired outcomes. Alternative metrological mapping of substantive variations in what is measured facilitate paradigmatically distinct formative approaches to management. These approaches are informed by clear, communicable, and reproducible evidence as to developmental sequences aligning current positions in relation to what has been accomplished, what should be saved for later, and what comes next (Black et al., 2011; Cardace et al., 2021; Dozier et al., 2023; Fisher, 2013; Morell et al., 2017).

6.7.1 The Basel Committee on Banking Supervision

The Basel Committee on Banking Supervision (BCBS) is an informal body with 45 international members that recommends policy solutions, common approaches, and risk measurement standards, such as those included in the Basel Accords, useful in solving problems of financial capacity and information (BCBS, 2011, 2021). The overall intention is to minimize the possibility of large-scale financial catastrophes. The BCBS and the banking industry at large seem unaware of the limits of traditional methods of financial accounting and statistical modeling as mechanisms for measuring and managing the wide range of substantive risks encountered in commercial enterprises.

Three quantified forms of risk (credit, operational, and market) are used in determining minimum capital requirements. Banks develop bespoke risk measurement systems to better align the real world economic definition of capital with its regulatory representation. Doing so works in the banks' favor to reduce their capital requirements. Other risks not considered fully quantifiable and not included in the maintenance of regulatory capital are systemic risk, pension risk, concentration risk, strategic risk, reputational risk, liquidity risk, and legal risk, which Basel II combines into the single category of residual risk subject to supervisory review. All of the forms of risk are taken into account in the Internal Capital Adequacy Assessment Process, which is expected to be an integral part of financial institutions' material business activities and decisions.

Basel II allowed credit institutions three possible methods for assessing operational risk: the Basic Indicator Approach, the Standard Approach, and the Advanced Measurement Approach. The latter is the most sophisticated, the most sensitive to operational risk, and the one that allows banks the most leeway in developing their own empirical risk models, though these models, and reversion to the less sophisticated

Basic Indicator or Standard Approaches, are subject to supervisory approval. Figini et al. (2010) propose an analytical integration of operational and financial risk assessments. In their conclusion, they suggest the possible relevance of another class of probabilistic models, those described by Rasch (1960). But the authors take an unnecessarily narrow view of these models when they say only an "ex-post approach" enabling observation of flaws and shortcomings in data can be supported, as though no implications for data design can be fed back to improve system quality. Their perspective is contingent on the unnecessary and counterproductive assumption that measurement is inherently a matter of statistical modeling and data analysis, and that it must accept whatever data are provided as definitive, on the basis of the authority of those providing it.

Overviews of measurement methods used in financial management and risk assessment (Chornous & Ursulenko, 2013; Kedarya et al., 2023) indicate that the field is dominated by statistical models, with little or no awareness of the availability of scientific models offering alternative capacities.

6.7.2 The United Nations Agenda 2030 and the Sustainable Development Goals

The United Nations' (UN) Sustainable Development Goals (SDGs) were launched in 2016 as part of the Agenda 2030, a global effort aimed at reducing poverty, promoting peace and prosperity, and protecting the environment. The SDGs encompass 17 goals that are intended to inspire collaborations across areas of governance, economic inequality, education, climate change, health care, and innovations in commercial sustainability, social justice, and reduced violence. UN Development Programme (UNDP) efforts in 170 countries address SDG partnerships with citizens, their governments, the private sector, and civil society at every level.

Evaluations of efforts aimed at realizing the SDG targets and goals are quantified in ways that unnecessarily and counterproductively alter how sustainable development is conceptualized and operationalized (Engebretsen et al., 2017; Merry, 2019; Ulbrich et al., 2019). The focus on numeric counts instead of on measured quantities restricts accountability to static accomplishments and cannot encompass needed processual concerns with systemic change. Profound disconnects separate the lofty ideals of sustainability mission, vision, and values statements from any hope for their broad scale fulfillment. Unexamined assumptions associated with the measurement and management of SDG policies and programs in education, for instance, "reconfigure education problems and issues in ways that invite certain possibilities for deliberation and intervention at the expense of others" (Grek, 2020, p. 163), with no means for altering the dynamic of the trade-offs involved. The co-production of scientific and social orders takes place in unintended and dysfunctional ways because of inattention to the consequences of methodical quantifications imposed irrespective of local needs.

Metrologically quality assured measurement, however, could map qualitatively meaningful variation in goal attainment across local circumstances in comparable ways capable of informing accountability and improvement efforts adaptable to the specific needs and challenges encountered (Fisher & Wilson, 2019, 2020; Fisher et al., 2019; Lips da Cruz et al., 2019). Alternative approaches to locally organized participatory social ecologies could proactively promote citizen participation in more organically relational, deliberate, and adaptive ways (Glock-Grueneich & Ross, 2008; Ross & Commons, 2008; Ross & Glock-Grueneich, 2008a/b; Morrison & Fisher, 2018–2024).

6.7.3 The Commission on the Measurement of Economic Performance and Social Progress

In early 2008, Joseph Stiglitz, Amartya Sen, and Jean Paul Fitoussi were involved in creating The Commission on the Measurement of Economic Performance and Social Progress (CMEPSP). Dissatisfactions with the conceptual limits of economic indicators prompted the formation of an expert panel exploring possibilities for new measurements of productivity and social progress. Stiglitz became President of the Commission, Sen an Advisor, and Fitoussi, the Coordinator. They convened an impressive group of economists affiliated with prestigious institutions in Europe and the USA, and they produced a 292-page report intended to open a discussion. As of March 2024, Google Scholar shows that the report has been cited over 7,700 times, indicating the report succeeded to a modest degree in realizing that goal.

The Commission's stated aims focus on how to present statistical information appropriately, while recognizing the importance of statistical indicators in "designing and assessing policies aiming at advancing the progress of society, as well as for assessing and influencing the functioning of economic markets." The Commission was organized into three groups addressing classical GDP issues, quality of life, and sustainability. Quality of life concerns current well-being, while sustainability focuses on capacities for continuing policies into the future. The report (p. 11) distinguishes natural, physical, human, and social stocks of capital as each needing to be considered in evaluating sustainability.

The report's sixth through tenth recommendations take up matters associated with the importance of incorporating both objective and subjective aspects of quality of life and well-being. The tenth recommendation (p. 16) urges statistical offices to raise "questions to capture people's life evaluations, hedonic experiences and priorities in their own survey." This is justified by the fact that "Research has shown that it is possible to collect meaningful and reliable data on subjective as well as objective well-being." Accordingly, "the types of question that have proved their value within small-scale and unofficial surveys should be included in larger-scale surveys undertaken by official statistical offices."

Almost 4,000 of the over 7,000 sources shown to cite the report in Google Scholar mention quality of life. This shows that the level of seriousness with which this recommendation has been received. Of those, 30 mention (but may not employ) an identified probabilistic model of individual-level measurement requiring separable parameters and sufficient statistics. Applications of such models could possibly lead to the calibration of instruments measuring in interval quantities. But none of those 30 articles mention metrology or the value of measurement defined in terms of instruments calibrated in defined units and distributed to end users in quality assured networks.

The foreword to Stiglitz, Sen, and Fitoussi's (2010, pp. ix–x) expansions on the CMEPSP report states:

> Our world, our society, and our economy have changed, and the measures have not kept pace.
>
> We will not change our behavior unless we change the ways we measure.
>
> A tremendous revolution awaits us – we can all feel it.
>
> This revolution will only be fully completed if it is first of all a revolution in our minds, in the way we think, in our mind sets and values.
>
> The problem stems from the fact that ultimately, without even realizing it, the statistics and the accounts were made to say things that they weren't saying and that they couldn't say. We have wound up mistaking our representations of wealth for the wealth itself, and our representations of reality for the reality itself. But reality always ends up having the last word.

But is a revolution in thinking a matter of individual will? Is educating and persuading individuals in the ways they must change sufficient to the task of bringing about the changes we want to see happen in how we measure and manage our world, our society, and our economy? Will changing the aggregate numbers reported and incorporated in broadly administered policies and procedures be able to change anything important? Why is no mention made of the possibility that improved measurement could empower individuals as entrepreneurs in new domains and markets commercializing human, social, and natural capital innovations? It seems that the investigations and recommendations of the CMEPSP report should be revisited in light of a metrologically informed perspective on the measurement of economic performance and social progress.

6.8 Closing thoughts

Given the limits imposed on thinking by the prevailing culture's available concepts, it should be obvious that thinking differently is not something individuals can simply decide to do. Merely changing the content of thinking without changing the infrastructural context informing it ought not count as thinking differently. Individuals certainly cannot each independently decide on new ways of thinking and coordinate their decisions and behaviors on the scale needed if we are to enact a revolution in our economic thinking and acting.

Changing our thinking in coordinated ways requires new conceptual resources, terms for representing them, and relational connections in the world. To change our thinking, we must change the terms of the languages we think in. The creation and distribution of those terms are inherently beyond the control of individuals. Metrology, however, offers means by which collectively projected structural invariances can be modeled, theoretically described, and systematically implemented in quality assured systems that remain open to local improvisations and continuous improvement. Metrology is a way of identifying individually relevant developmental paths of least resistance and narrating stories meaningful to all, and of making those paths and stories universally accessible, even though they never apply perfectly to anyone.

Foremost among the collectively constituted languages we think in (Bernstein, 2004; Fisher, 2012c, 2020, 2023a) are

1. the metrological infrastructures informing the conduct of science,
2. the legal terms of property rights and contract law,
3. the financial terms of markets and accounting, and
4. the communications of digital networks.

In the spirit of Prigogine and Stengers' (1984, p. 22) efforts, we can connect these domains by translating individuals' and groups' somewhat divergent and somewhat convergent interests, doing so in ways that overcome their isolation from one another and from the Earth, and open new channels of communication between science and society.

In today's GDP-driven economies and one-size-fits-all appeals to the law of averages, we have indeed confused "our representations of reality for the reality itself" (Stiglitz et al., 2010, p. ix). But while there is no direct access to reality, no primordial given presence that serves as the ground we metaphorically stand on, we need no longer assume there is no alternative to centrally planned "statisticism," as Duncan (1984b, p. 226) called the fixation on centrally planned and executed numeric analyses instead of on metrologically traceable, quality assured, distributed networks of substantive measurements. Indeed, his emphasis on the lessons to be learned from historical metrology remains as valid now as it was 40 years ago.

Perhaps the juxtaposed contrasts of the measurement roads taken and not taken in the contexts of the Basel Accords, the UN Agenda 2030's Sustainable Development Goals, and Commission on the Measurement of Economic Performance and Social Progress will provoke some to consider new metrological horizons in future efforts of these kinds.

To summarize, the successes of modern science have altered the global environment to the point that its methods no longer work to advance quality of life and the economy. Modern thinking and methods are now part of the problem. Where modernism's isolated autonomous individuals are alienated from an indifferent independent reality they can only describe, an unmodern perspective recognizes and operationalizes a participatory integration of individual thinking with linguistic and metric standards embedded in the external environment. Taking the integrated organism-environment as the focal unit of natural selection, unmodern thinking alters the sociocognitive environment so as to amplify and feed back meaningful structural invariances – collectively projected coherent constructs – metrologically, informing individual thinking and communications in a shared social reality. Rich logical, aesthetic, and ethical potentials follow from extending to human, social, and natural capital the new institutional economics' insights into how the successes of capitalism have been contingent on the integration of metrology with legal property rights, capital markets, and communications networks (Fisher, 2012c, 2020, 2023a). Metrologically combining ideals, standards, and local implementations in pragmatic methods may galvanize a shared vision of the future humanity urgently needs if it is to imagine and create drastically transformed economies capable of preventing an ever-increasing array of impending human, social, and environmental catastrophes.

References

Abram, D. (1996). *The spell of the sensuous: Perception and language in a more-than-human world*. Vintage Books.

Ackermann, J. R. (1985). *Data, instruments, and theory: A dialectical approach to understanding science*. Princeton University Press.

Acs, Z. J., Estrin, S., Mickiewicz, T., & Szerb, L. (2018). Entrepreneurship, institutional economics, and economic growth: An ecosystem perspective. *Small Business Economics*, *51*(2), 501–514. https://doi.org/10.1007/s11187-018-0013-9

Akrich, M., Callon, M., & Latour, B. (2002). The key to success in innovation Parts I & II. *International Journal of Innovation Management*, *6*(2), 187–225.

Akrich, M., & Latour, B. (1992). A summary of a convenient vocabulary for the semiotics of human and nonhuman assemblies. In W. E. Bijker & J. J. Law (Eds.). *Shaping Technology/Building Society*. MIT Press.

Alagumalai, S., Durtis, D. D., & Hungi, N. (2005). *Applied Rasch measurement: A book of exemplars*. Springer-Kluwer.

Alder, K. (2002). *The measure of all things: The seven-year odyssey and hidden error that transformed the world*. The Free Press.

Alker, H. R. (1969). A typology of ecological fallacies. In M. Dogan & S. Rokkan (Eds.). *Quantitative ecological analysis in the social sciences* (pp. 69–86). MIT Press.

Allen, D. D., & Pak, S. (2023). Improving clinical practice with person-centered outcome measurement. In W. P. Fisher & S. J. Cano (Eds.). *Person centered outcome metrology* (pp. 53–105). Springer.

Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, *65*(331), 1248–1255.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*(1), 69–81.

Andersen, E. B. (1999). Sufficient statistics in educational measurement. In G. N. Masters & J. P. Keeves (Eds.). *Advances in measurement in educational research and assessment* (pp. 122–125). Pergamon.

Andersson, P. (2015). Scaffolding of task complexity awareness and its impact on actions and learning. *Action Learning and Action Research Journal*, *21*(1), 124–147.

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*, 449–460.

Andrich, D. (1988). *Sage University Paper Series on Quantitative Applications in the Social Sciences. Vol. series no. 07-068: Rasch models for measurement*. Sage Publications.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J. A. Keats, R. Taft, R. A. Heath, & S. H. Lovibond (Eds.). *Mathematical and Theoretical Systems: Proceedings of the 24th International Congress of Psychology of the International Union of Psychological Science, Vol.* 4 (pp. 7–16). Elsevier Science Publishers.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*(1), I-7–I-16.

Andrich, D. (2005). The Rasch model explained. In S. Alagumalai, D. D. Durtis, & N. Hungi (Eds.). *Applied Rasch measurement: A book of exemplars* (pp. 308–328). Springer-Kluwer.

Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika*, *75*(2), 292–308.

Arias-Maldonado, M. (2015). *Environment and society: Socionatural relations in the Anthropocene*. Cham: Springer.

Ashworth, W. J. (2004, November 19). Metrology and the state: Science, revenue, and commerce. *Science*, *306*(5700), 1314–1317.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554.

Ballard, E. G. (1978). *Man and technology: Toward the measurement of a culture*. Duquesne University Press.

Bamber, D., & van Santen, J. P. H. (1985). How many parameters can a model have and still be testable? *Journal of Mathematical Psychology*, *29*, 443–473.

Banks, E. (2004). The philosophical roots of Ernst Mach's economy of thought. *Synthese*, *139*(1), 23–53.

Barab, S. A., & Plucker, J. A. (2002). Smart people or smart contexts? Cognition, ability, and talent development in an age of situated approaches to knowing and learning. *Educational Psychologist*, *37*(3), 165–182.

Barndorff-Nielsen, O. (1973). *Exponential families and conditioning*. Aarhus Universite.

Barney, M., & Barney, F. (2024). Transdisciplinary measurement through AI: Hybrid metrology and psychometrics powered by Large Language Models. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI* (p. in press). De Gruyter.

Barney, M., & Fisher, W. P., Jr. (2016). Adaptive measurement and assessment. *Annual Review of Organizational Psychology and Organizational Behavior*, *3*, 469–490. https://www.annualreviews.org/doi/abs/10.1146/annurev-orgpsych-041015-062329

Barney, M., & Fisher, W. P., Jr. (2017). Avoiding AI Armageddon with metrologically-oriented psychometrics. *18th International Congress of Metrology*, *09005*, 1–6. https://doi.org/10.1051/metrology/201709005

Bartolucci, F., Pennoni, F., & Vittadini, G. (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioral Statistics*, *36*(4), 491–522.

Barzel, Y. (1982). Measurement costs and the organization of markets. *Journal of Law and Economics*, *25*, 27–48.

Bateson, G. (1972). *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology*. University of Chicago Press.

Bateson, G. (1978). Number is different from quantity. *CoEvolution Quarterly*, 17, 44–46 [Reprinted from pp. 53–58 in Bateson, G. (1979). Mind and Nature: A Necessary Unity. New York: E. P. Dutton.]. http://www.wholeearth.com/issue/2017/article/295/number.is.different.from.quantity

Bateson, G. (1991). *A sacred unity: Further steps to an ecology of mind*. R. E. Donaldson, (Ed.). HarperOne.

Bateson, G., Jackson, D. D., Haley, J., & Weakland, J. (1956). Toward a theory of schizophrenia. *Behavioral Science*, *1*(4), 251–264.

BCBS: Basel Committee on Banking Supervision. (2011). *Operational risk -- supervisory guidelines for the advanced measurement approaches*. Basel, Switzerland:. Bank for International Settlements. https://www.bis.org/publ/bcbs196.pdf (63 pp.)

BCBS: Basel Committee on Banking Supervision. (2021). *Climate-related financial risks - measurement methodologies*. Basel, Switzerland:. Bank for International Settlements. (56 pp.)

Belvedere, S. L., & de Morton, N. A. (2010). Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. *Journal of Clinical Epidemiology*, *63*(12), 1287–1297.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olson-Buchanan (Eds.). *Innovations in computerized assessment* (pp. 67–91). Lawrence Erlbaum Associates, Inc., Publishers.

Bernstein, W. J. (2004). *The birth of plenty: How the prosperity of the modern world was created*. McGraw-Hill.

Berti, M., & Simpson, A. V. (2021). The dark side of organizational paradoxes: The dynamics of disempowerment. *Academy of Management Review*, *46*(2), 252–274.

Bettelheim, B. (1967). *The empty fortress: Infantile autism and the birth of the self*. The Free Press.

Bezruczko, N. (Ed.). (2005). *Rasch measurement in health sciences*. JAM Press.

Bizouarn, P. (2016). L'éco-épidémiologie: Vers une épidémiologie de la complexité. *Médicine/Sciences*, *32*(5), 500–505.

Björgvinsson, E. (2014). The making of cultural commons: Nasty old film distribution and funding. In P. Ehn, E. M. Nilsson, & R. Topgaard (Eds.). *Making futures: Marginal notes on innovation, design, and democracy* (pp. 187–225). MIT Press.

Black, M. (1962). *Models and metaphors*. Cornell University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–74.

Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives*, *9*, 1–52.

Blok, A., Farias, I., & Roberts, C. (Eds.). (2020). *The Routledge companion to Actor-Network Theory*. Routledge.

Bohm, D. (1980). *Wholeness and the implicate order*. Routledge & Kegan Paul.

Bohm, D. (1986). A new theory of the relationship of mind and matter. *The Journal of the American Society for Psychical Research*, *80*(2), 113–135.

Bohm, D., & Hiley, B. J. (1984). Measurement understood through the quantum potential approach. *Foundations of Physics*, *14*(3), 255–274.

Bohm, D., Hiley, B. J., & Kaloyerou, P. N. (1987). An ontological basis for the quantum theory. *Physics Reports*, *144*(6), 321–375.

Bohr, N. (1963). *Essays 1958–1962 on atomic physics and human knowledge*. John Wiley & Sons.

Boundas, C. V. (Ed.). (2018). *Schizoanalysis and ecosophy: Reading Deleuze and Guattari*. Bloomsbury Publishing.

Bowker, G. C. (2015). Susan Leigh Star Special Issue. *Mind, Culture, and Activity*, *22*(2), 89–91.

Bowker, G., Star, S. L., Gasser, L., & Turner, W. (Eds.). (2014). *Social science, technical systems, and cooperative work: Beyond the great divide*. Psychology Press.

Bowker, G., Timmermans, S., Clarke, A. E., & Balka, E. (Eds.). (2015). *Boundary objects and beyond: Working with Leigh Star*. MIT Press.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.). *Robustness in statistics* (pp. 201–235). Academic Press, Inc.

Boyer, C. B. (1949). *The history of the calculus and it conceptual development*. Dover.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of pair comparisons. *Biometrika*, *63*, 324–345.

Brier, S. (2010). Cybersemiotics: An evolutionary world view going beyond entropy and information into the question of meaning. *Entropy*, *12*(8), 1902–1920.

Brier, S. (2011). Ethology and the Sebeokian way from zoosemiotics to cyber(bio)semiotics. In P. Cobley, J. Deely, K. Kull, & S. Petrilli (Eds.). *Semiotics continues to astonish: Thomas A. Sebeok and the doctrine of signs* (pp. 41–84). De Gruyter.

Brier, S. (2013). Cybersemiotics: A new foundation for transdisciplinary theory of information, cognition, meaningful communication and the interaction between nature and culture. *Integral Review*, *9*(2), 220–263.

Brier, S. (2021). Cybersemiotic systemic and semiotical based transdisciplinarity. In C. Vidales & S. Brier (Eds.). *Introduction to cybersemiotics: A transdisciplinary perspective* (pp. 17–31). Springer.

Brink, N. E. (1972). Rasch's logistic model vs the Guttman model. *Educational and Psychological Measurement*, *32*, 921–927.

Brogden, H. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika*, *42*, 631–634.

Brown, T. (2003). *Making truth: Metaphor in science*. University of Illinois Press.

Brown, T. (2008). Design thinking. *Harvard Business Review*, *86*(6), 84–92.

Brown, J. C., Goldstein, J. E., Chan, T. L., Massof, R., & Ramulu, P., & Low Vision Research Network Study Group. (2014). Characterizing functional complaints in patients seeking outpatient low-vision services in the United States. *Ophthalmology*, *121*(8), 1655–1662.

Bundgaard, P. F. (2019). The structure of our concepts: A critical assessment of Conceptual Metaphor Theory as a theory of concepts. *Cognitive Semiotics*, *12*(1), 1–11.

Burtt, E. A. (1954). *The metaphysical foundations of modern physical science* (Rev. ed.), Doubleday Anchor.

Butterfield, H. (1957). *The origins of modern science (Rev. ed.)*. The Free Press.

Cajori, F. (1999). *A history of mathematics, 5th ed*. AMS Chelsea Publishing Co.

Callon, M. (1995). Four models for the dynamics of science. In S. Jasanoff, G. E. Markle, J. C. Petersen, & T. Pinch (Eds.). *Handbook of science and technology studies* (pp. 29–63). Sage Publications.

Cano, S., Klassen, A. F., & Pusic, A. L. (2009). The science behind quality-of-life measurement: A primer for plastic surgeons. *Plastic and Reconstructive Surgery*, *123*(3), 98e–106e.

Cano, S., Melin, J., Fisher, W. P., Jr, Stenner, A. J., & Pendrill, L., & EMPIR NeuroMet 15HLT04 Consortium. (2018). Patient-centred cognition metrology. *Journal of Physics: Conference Series*, *1065*, 072033. https://iopscience.iop.org/article/10.1088/1742-6596/1065/7/072033/meta

Cano, S., Pendrill, L., Melin, J., & Fisher, W. P., Jr. (2019). Towards consensus measurement standards for patient-centered outcomes. *Measurement*, *141*, 62–69. https://doi.org/10.1016/j.measurement.2019.03.056

Capra, F., & Luisi, P. L. (2014). *The systems view of life: A unifying vision*. Cambridge University Press.

Caraman, S., & Caraman, L. (2021). Metaphor: A key element of beauty in poetry and mathematics. In B. Sriraman (Ed.). *Handbook of the mathematics of the arts and sciences* (pp. 1015–1044). Springer International Publishing.

Cardace, A., Wilson, M., & Metz, K. E. (2021). Designing a learning progression about micro-evolution to inform instruction and assessment in elementary science. *Education Sciences*, *11*(10), 609. https://doi.org/10.3390/educsci11100609

Cartwright, N. (1983). *How the laws of physics lie*. Oxford University Press.

Chaitin, G. J. (1994). Randomness and complexity in pure mathematics. *International Journal of Bifurcation and Chaos*, *4*(1), 3–15. http://www.worldscientific.com/doi/pdf/10.1142/S0218127494000022

Chandrasekhar, S. (1979). Beauty and the quest for beauty in science. *Physics Today*, *32*(7), 25–30.

Chemero, A. (2013). Radical embodied cognitive science. *Review of General Psychology*, *17*(2), 145–150.

Chien, T. W., Chang, Y., Wen, K. S., & Uen, Y. H. (2018). Using graphical representations to enhance the quality-of-care for colorectal cancer patients. *European Journal of Cancer Care*, *27*(1), e12591.

Chornous, G., & Ursulenko, G. (2013). Risk management in banks: New approaches to risk assessment and information support. *Ekonomika*, *92*(1), 120–132.

Christensen, K. B., Kreiner, S., & Mesbah, M. (2013). *Rasch models in health*. John Wiley & Sons, Inc.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*, 186–190.

Cliff, N. (1993). What is and isn't measurement. In G. Keren & C. Lewis (Eds.). *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 59–93). Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round (p < 0.05). *American Psychologist*, *49*, 997–1003. https://psycnet.apa.org/record/1995-12080-001

Cohen, T. (1979). Metaphor and the cultivation of intimacy. In S. Sacks (Ed.). *On metaphor* (pp. 1–10). University of Chicago Press.

Cole, K. C. (1985). *Sympathetic vibrations: Reflections on physics as a way of life*. Bantam Books.

Commons, M. L., & Bresette, L. M. (2006). Illuminating major creative scientific innovators with postformal stages. In C. Hoare (Ed.). *Handbook of adult development and learning* (pp. 255–280). Oxford University Press.

Commons, M. L., & Duong, T. Q. (2019). Understanding terrorism: A behavioral developmental approach. *Ethics, Medicine and Public Health*, *8*, 74–96.

Commons, M. L., Gane-McCalla, R., Barker, C. D., & Li, E. Y. (2014). The model of hierarchical complexity as a measurement system. *Behavioral Development Bulletin*, *19*(3), 9–14.

Commons, M. L., & Goodheart, E. A. (2007). Consider stages of development in preventing terrorism: Does government building fail and terrorism result when developmental stages of governance are skipped? *Journal of Adult Development*, *14*, 91–111.

Commons, M. L., & Goodheart, E. A. (2008). Cultural progress is the result of developmental level of support. *World Futures: The Journal of New Paradigm Research*, *64*(5–7), 406–415. https://doi.org/10.1080/02604020802301360

Commons, M. L., Goodheart, E. A., Pekker, A., Dawson-Tunik, T. L., Draney, K., & Adams, K. M. (2008). Using Rasch scaled stage scores to validate orders of hierarchical complexity of balance beam task sequences. *Journal of Applied Measurement*, *9*, 182–199.

Commons, M. L., & Richards, F. A. (2002). Four postformal stages. In J. Demick & C. Andreoletti (Eds.). *Handbook of adult development* (pp. 199–219). Plenum Press.

Commons, M. L., & Ross, S. N. (Eds.). (2008). Post-formal thought and hierarchical complexity [Special issue]. *World Futures: Journal of General Evolution*, 64(5–7). doi:10.1080/02604020802301121

Commons, M. L., Ross, S. N., & Bresette, L. M. (2011). The connection between postformal thought, stage transition, persistence, and ambition and major scientific innovations. In D. Artistico, J. Berry, J. Black, D. Cervone, C. Lee, & H. Orom (Eds.). *The Oxford handbook of reciprocal adult development and learning* (pp. 287–301). Oxford University Press.

Confrey, J., Shah, M., & Toutkoushian, E. (2021). Validation of a learning trajectory-based diagnostic mathematics assessment system as a trading zone. *Frontiers in Education: Assessment, Testing and Applied Measurement*, *6*(654353). doi:10.3389/feduc.2021.654353

Connolly, A. J., Nachtman, W., & Pritchett, E. M. (1971/2007). *Keymath: Diagnostic Arithmetic Test*. American Guidance Service. https://images.pearsonclinical.com/images/pa/products/keymath3_da/km3-da-pub-summary.pdf

Cook, T. A. (1914/1979). *The curves of life*. Dover.

Cozzo, E., Baños, R. A., Meloni, S., & Moreno, Y. (2013). Contact-based social contagion in multiplex networks. *Physical Review*, *88*, 050801.

Danesi, M. (2017). Semiotics as a metalanguage for the sciences. In K. Bankov & P. Cobley (Eds.). *Semiotics and its masters* (pp. 61–81). De Gruyter.

Danilov, I. V., & Mihailova, S. (2021). Neuronal coherence agent for shared intentionality: A hypothesis of neurobiological processes occurring during social interaction. *OBM Neurobiology*, *5*(4), 1. doi:10.21926/obm.neurobiol.2104113

Dawson, T. L. (2003). A stage is a stage is a stage: A direct comparison of two scoring systems. *Journal of Genetic Psychology*, *164*, 335–364.

Dawson, T. L. (2004). Assessing intellectual development: Three approaches, one sequence. *Journal of Adult Development*, *11*(2), 71–85.

Dawson-Tunik, T. L., Commons, M., Wilson, M., & Fischer, K. (2005). The shape of development. *The European Journal of Developmental Psychology*, *2*, 163–196.

Dear, P. (1992). From truth to disinterestedness in the seventeenth century. *Social Studies of Science*, *22*, 619–631.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. (Statistics for Social and Behavioral Sciences). Springer-Verlag.

de Castro, C., & McShea, D. W. (2022). Applying the Prigogine view of dissipative systems to the major transitions in evolution. *Paleobiology*, *48*(4), 711–728.

Deleuze, G., & Guattari, F. (1977). *Anti-Oedipus: Capitalism and schizophrenia*. Penguin.

Derrida, J. (1976). *Of grammatology*. G. C. Spivak (Trans.). Johns Hopkins University Press.

Derrida, J. (2003). Interview on writing. In G. A. Olson & L. Worsham (Eds.), *Critical intellectuals on writing* (pp. 61–69). State University of New York Press.

de Solla Price, D. J. (1986). Of sealing wax and string. In *Little science, big science–and beyond* (pp. 237–253). Columbia University Press.

De Soto, H. (2000). *The mystery of capital: Why capitalism triumphs in the West and fails everywhere else*. Basic Books.

Dewey, J. (1954). *The public and its problems*. Swallow Press, Ohio University Press.

Dewey, J. (2012). *Unmodern philosophy and modern philosophy*. P. Deen, (Ed.). Southern Illinois University Press.

Dirac, P. A. M. (1963). The evolution of the physicist's picture of nature. *Scientific American*, *208*(5), 45–53.

Dosch, H. G., Müller, V. F., & Sieroka, N. (2006). *Quantum field theory in a semiotic perspective*. Springer Science & Business Media.

Dozier, S. J., MacPherson, A., Morell, L., Gochyyev, P., & Wilson, M. (2023). A learning progression for understanding interdependent relationships in ecosystems. *Sustainability*, *15*(19), 14212. https://doi.org/10.3390/su151914212

Duncan, O. D. (1984a). Measurement and structure: Strategies for the design and analysis of subjective survey data. In C. F. Turner & E. Martin (Eds.). *Surveying subjective phenomena, Volume* 1 (pp. 179–229). Russell Sage Foundation.

Duncan, O. D. (1984b). *Notes on social measurement: Historical and critical*. Russell Sage Foundation.

Duncan, O. D. (1984c). Rasch measurement: Further examples and discussion. In C. F. Turner & E. Martin (Eds.). *Surveying subjective phenomena, Volume 2* (pp. 367–403). Russell Sage Foundation.

Duncan, O. D. (1992). What if? *Contemporary Sociology*, *21*(5), 667–668.

Duncan, O. D., & Stenbeck, M. (1988). Panels and cohorts: Design and model in the study of voting turnout. In C. C. Clogg (Ed.). *Sociological Methodology 1988* (pp. 1–35). American Sociological Association.

Durham, I., & Rickles, D. (2017). *Information and interaction: Eddington, Wheeler, and the limits of knowledge*. Springer

Einstein, A. (1954). *Ideas and opinions*. Bonanza Books.

Ekstig, B. (2010). Biological and cultural evolution in a common universal trend of increasing complexity. *World Futures*, *66*(6), 435–448.

Elbaum, B., Fisher, W. P., Jr, & Coulter, W. A. (2011). Measuring schools' efforts to partner with parents of children served under IDEA: Scaling and standard setting for accountability reporting. *Journal of Applied Measurement*, *12*(3), 261–278.

Elnegahy, S., Jin, H., & Kim, H. (2022). Fairness, justice, and language assessment. *Language Assessment Quarterly*, *19*(1), 102–105. doi:10.1080/15434303.2021.1922412

Elsbach, K. D., Barr, P. S., & Hargadon, A. B. (2005). Identifying situated cognition in organizations. *Organization Science*, *16*(4), 422–433.

Embretson, S. E. (1996). Item Response Theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*(3), 201–212.

Embretson, S. E. (2010). *Measuring psychological constructs: Advances in model-based approaches*. American Psychological Association.

Engebretsen, E., Heggen, K., & Ottersen, O. P. (2017). The Sustainable Development Goals: Ambiguities of accountability. *The Lancet*, *389*(10067), 365.

Estep, M. L. (1993). On the law of excluded middle and arguments for multivalued logics in systems inquiry. *Cybernetics and Systems*, *24*(3), 243–254.

Falmagne, J.-C., & Doignon, J.-P. (1997). Stochastic evolution of rationality. *Theory and Decision*, *43*, 107–138.

Favela, L. H. (2020). Cognitive science as complexity science. *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(4), e1525.

Feynman, R. (1965). *The character of physical law*. MIT Press.

Figini, S., Kenett, R. S., & Salini, S. (2010). *Integrating operational and financial risk assessments* (Dipartimento Di Scienze Economiche Aziendali e Statistiche No. 2010-02). Milano, Italy: Universita Degli Studi di Milano.

Finkelstein, L. (1975). Representation by symbol systems as an extension of the concept of measurement. *Kybernetes*, *4*(4), 215–223.

Fischer, G. H. (1968). *Psychologische testtheorie*. Huber.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, *46*(1), 59–77.

Fischer, G. H., & Molenaar, I. (1995). *Rasch models: Foundations, recent developments, and applications*. Springer-Verlag.

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*, *87*, 477–531.

Fischer, K. W., & Dawson, T. L. (2002). A new kind of developmental science: Using models to integrate theory and research: Comment. *Monographs of the Society for Research in Child Development*, *67*(1), 156–167.

Fischer, K. W., & Farrar, M. J. (1987). Generalizations about generalization: How a theory of skill development explains both generality and specificity. *International Journal of Psychology*, *22*(5–6), 643–677.

Fisher, A. G. (1997). Multifaceted measurement of daily life task performance: Conceptualizing a test of instrumental ADL and validating the addition of personal ADL tasks. *Physical Medicine & Rehabilitation State of the Art Reviews: Outcome Measurement*, *11*(2), 289–303.

Fisher, C. M., Demir-Caliskan, O., Yingying Hua, M., & Cronin, M. A. (2021). Trying not to try: The paradox of intentionality in jazz improvisation and its implications for organizational scholarship. In R. Bednarek, M. Pina E Cunha, J. Schad, & W. K. Smith (Eds.). *Research in the Sociology of Organizations: Vol. 73b. Interdisciplinary Dialogues on Organizational Paradox: Investigating Social Structures and Human Expression, Part B* (pp. 123–137). Emerald Publishing Limited.

Fisher, F. M. (1959). Generalization of the rank and order conditions for identifiability. *Econometrica*, *27*, 431–447.

Fisher, I. (1930). *The theory of interest*. Macmillan.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A*, *222*, 309–368.

Fisher, R. A. (1934a). Indeterminism and natural selection. *Philosophy of Science*, *1*(1), 99–117. https://www.cambridge.org/core/journals/philosophy-of-science/article/abs/indeterminism-and-natural-selection/CD152AD8CFEE525A97749033D3518460#

Fisher, R. A. (1934b). Two new properties of mathematical likelihood. *Proceedings of the Royal Society, A*, *144*, 285–307.

Fisher, R. A. (1954). Retrospect of the criticisms of the theory of natural selection. In J. Huxley, A. C. Hardy, & E. B. Ford (Eds.). *Evolution as a process* (pp. 84–98). George Allen & Unwin Ltd.

Fisher, W. P., Jr. (1988). *Truth, method, and measurement: The hermeneutic of instrumentation and the Rasch model* [Diss]. Dissertation Abstracts International (Dept. of Education, Division of the Social Sciences), 49, 0778A. (376 pages, 23 figures, 31 tables)

Fisher, W. P., Jr. (1990). Mangoes, metaphors, and a measure of meaning. Presented at the African Studies Association, Baltimore, MD, November.

Fisher, W. P., Jr. (1992a). Objectivity in measurement: A philosophical history of Rasch's separability theorem. In M. Wilson (Ed.). *Objective measurement: Theory into practice. Vol. I* (pp. 29–58). Ablex Publishing Corporation.

Fisher, W. P., Jr. (1992b). Stochastic resonance and Rasch measurement. *Rasch Measurement Transactions*, *5*(4), 186–187 http://www.rasch.org/rmt/rmt54k.htm.

Fisher, W. P., Jr. (1995). Metaphor as virtual measurement, and Quantitative post-structuralist performance assessment. *Two papers presented at the Eighth International Objective Measurement Workshop*, Tolman Hall, University of California, Berkeley, April 16.

Fisher, W. P., Jr. (1997a). Physical disability construct convergence across instruments: Towards a universal metric. *Journal of Outcome Measurement*, *1*(2), 87–113. http://jampress.org/JOM_V1N2.pdf

Fisher, W. P., Jr. (1997b). What scale-free measurement means to health outcomes research. *Physical Medicine & Rehabilitation State of the Art Reviews*, *11*(2), 357–373.

Fisher, W. P., Jr. (1999). Foundations for health status metrology: The stability of MOS SF-36 PF-10 calibrations across samples. *Journal of the Louisiana State Medical Society*, *151*(11), 566–578. https://europepmc.org/article/med/10618861

Fisher, W. P., Jr. (2000). Objectivity in psychosocial measurement: What, why, how. *Journal of Outcome Measurement*, *4*(2), 527–563. http://jampress.org/JOM_V4N2.pdf

Fisher, W. P., Jr. (2002). "The Mystery of Capital" and the human sciences. *Rasch Measurement Transactions*, *15*(4), 854 [http://www.rasch.org/rmt/rmt154j.htm].

Fisher, W. P., Jr. (2003). Mathematics, measurement, metaphor, metaphysics: Parts I & II. *Theory & Psychology*, *13*(6), 753–828.

Fisher, W. P., Jr. (2004). Meaning and method in the social sciences. *Human Studies: A Journal for Philosophy and the Social Sciences*, *27*(4), 429–454.

Fisher, W. P., Jr. (2007). Living capital metrics. *Rasch Measurement Transactions*, *21*(1), 1092–1093 [http://www.rasch.org/rmt/rmt211.pdf].

Fisher, W. P., Jr. (2009a). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement*, *42*(9), 1278–1287.

Fisher, W. P., Jr. (2009b). *NIST Critical national need idea White Paper: metrological infrastructure for human, social, and natural capital* (Tech. Rep. http://www.nist.gov/tip/wp/pswp/upload/202_metrological_infrastructure_for_human_social_natural.pdf). Washington, DC:. National Institute for Standards and Technology. (11 pages)

Fisher, W. P., Jr. (2010a). *Measurement, reduced transaction costs, and the ethics of efficient markets for human, social, and natural capital*, Bridge to Business Postdoctoral Certification, Freeman School of Business, Tulane University, New Orleans, Louisiana (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2340674).

Fisher, W. P., Jr. (2010b). Reducible or irreducible? Mathematical reasoning and the ontological method. In M. Garner, G. Engelhard Jr., W. P. Fisher Jr, & M. Wilson (Eds.). *Advances in Rasch measurement, Vol.* 1 (pp. 12–44). JAM Press.

Fisher, W. P., Jr. (2010c, January 13). *Reinventing capitalism: Diagramming living capital flows in a green, sustainable, and responsible economy*. LivingCapitalMetrics.com blog: https://livingcapitalmetrics.wordpress.com/2010/01/13/reinventing-capitalism/.

Fisher, W. P., Jr. (2010d). The standard model in the history of the natural sciences, econometrics, and the social sciences. *Journal of Physics Conference Series*, *238*(1), http://iopscience.iop.org/1742-6596/238/1/012016/pdf/1742-6596_238_1_012016.pdf.

Fisher, W. P., Jr. (2011a). Bringing human, social, and natural capital to life: Practical consequences and opportunities. In N. Brown, B. Duckor, K. Draney, & M. Wilson (Eds.). *Advances in Rasch Measurement, Vol.* 2 (pp. 1–27). JAM Press.

Fisher, W. P., Jr. (2011b). Metaphor as measurement, and vice versa: Convergence and separation of figure and meaning in a Mawri proverb [Modified version of a paper presented to the African Studies Association, 1990]. Social Science Research Network. Retrieved 7 April 2014, from http://ssrn.com/abstract=1747967

Fisher, W. P., Jr. (2011c). Stochastic and historical resonances of the unit in physics and psychometrics. *Measurement: Interdisciplinary Research and Perspectives*, *9*, 46–50.

Fisher, W. P., Jr. (2012a, July 12). Metaphor as measurement and vice versa: Love is a rose. In A. Maul (Chair), *Metaphors and measurement: An invited symposium on validity*. Lincoln, Nebraska: International Meeting of the Psychometric Society. http://www.slideshare.net/wpfisherjr/fisher-imps2012c-metaphor

Fisher, W. P., Jr. (2012b). Measure and manage: Intangible assets metric standards for sustainability. In J. Marques, S. Dhiman, & S. Holt (Eds.). *Business administration education: Changes in management and leadership strategies* (pp. 43–63). Palgrave Macmillan.

Fisher, W. P., Jr. (2012c). What the world needs now: A bold plan for new standards. *Standards Engineering*, *64*(3), 1 & 3–5. http://ssrn.com/abstract=2083975.

Fisher, W. P., Jr. (2013). Imagining education tailored to assessment as, for, and of learning: Theory, standards, and quality improvement. *Assessment and Learning*, *2*, 6–22.

Fisher, W. P., Jr. (2017a). A practical approach to modeling complex adaptive flows in psychology and social science. *Procedia Computer Science*, *114*, 165–174. https://doi.org/10.1016/j.procs.2017.09.027

Fisher, W. P., Jr. (2017b). Provoking professional identity development: The legacy of Benjamin Drake Wright. In M. Wilson & W. P. Fisher Jr. (Eds.). *Psychological and social measurement: The career and contributions of Benjamin D. Wright* (pp. 135–162). Springer.

Fisher, W. P., Jr. (2019). A nondualist social ethic: Fusing subject and object horizons in measurement. *TMQ–Techniques, Methodologies, and Quality, 10*(Special Issue on Health Metrology), 21–40.

Fisher, W. P., Jr. (2020). Contextualizing sustainable development metric standards: Imagining new entrepreneurial possibilities. *Sustainability*, *12*(9661), 1–22. https://doi.org/10.3390/su12229661

Fisher, W. P., Jr. (2021). Bateson and Wright on number and quantity: How to not separate thinking from its relational context. *Symmetry, 13*(1415). https://doi.org/10.3390/sym13081415

Fisher, W. P., Jr. (2022a). Aiming higher in conceptualizing manageable measures in production research. In N. Durakbasa & M. G. Gençyilmaz (Eds.). *Digitizing production systems: Selected papers from ISPR2021, October 07–09, 2021 online, Turkey* (pp. xix–xxxix). Springer Verlag. https://link.springer.com/content/pdf/bfm%3A978-3-030-90421-0%2F1

Fisher, W. P., Jr. (2022b). Contrasting roles of measurement knowledge systems in confounding or creating sustainable change. *Acta IMEKO*, *11*(4), 1–7. https://acta.imeko.org/index.php/acta-imeko/article/view/1330

Fisher, W. P., Jr. (2023a). Measurement systems, brilliant results, and brilliant processes in healthcare: Untapped potentials of person-centered outcome metrology for cultivating trust. In W. P. Fisher Jr. & S. Cano (Eds.). *Person-centered outcome metrology: Principles and applications for high stakes decision making* (pp. 357–396). Springer. https://link.springer.com/book/10.1007/978-3-031-07465-3

Fisher, W. P., Jr. (2023b). Separation theorems in econometrics and psychometrics: Rasch, Frisch, two Fishers, and implications for measurement. *Journal of Interdisciplinary Economics*, *35*(1), 29–60. https://journals.sagepub.com/doi/10.1177/02601079211033475

Fisher, W. P., Jr. (2024). Entropy, deterministic chaos, and new forms of intelligibility: A shared frame of reference for physics and psychology. *International Journal of Advances in Production Research, 1*(1), 46–83. https://dergipark.org.tr/en/pub/ijapr/issue/83619/1453834.

Fisher, W. P., Jr, & Burton, E. (2010). Embedding measurement within existing computerized data systems: Scaling clinical laboratory and medical records heart failure data to predict ICU admission. *Journal of Applied Measurement*, *11*(2), 271–287.

Fisher, W. P., Jr., & Cano, S. (Eds.). (2023). *Person-centered outcome metrology: Principles and applications for high stakes decision making*. (Springer Series in Measurement Science & Technology). Springer. https://link.springer.com/book/10.1007/978-3-031-07465-3

Fisher, W. P., Jr, Elbaum, B., & Coulter, W. A. (2010). Reliability, precision, and measurement in the context of data from ability tests, surveys, and assessments. *Journal of Physics*: *Conference Series, 238*(012036), http://iopscience.iop.org/1742-6596/238/1/012036/pdf/1742-6596_238_1_012036.pdf.

Fisher, W. P., Jr, Elbaum, B., & Coulter, W. A. (2012). Construction and validation of two parent-report scales for the evaluation of early intervention programs. *Journal of Applied Measurement*, *13*(1), 57–76.

Fisher, W. P., Jr, Eubanks, R. L., & Marier, R. L. (1997). Equating the MOS SF36 and the LSU HSI physical functioning scales. *Journal of Outcome Measurement*, *1*(4), 329–362. http://jampress.org/JOM_V1N4.pdf

Fisher, W. P., Jr, Harvey, R. F., & Kilgore, K. M. (1995a). New developments in functional assessment: Probabilistic models for gold standards. *NeuroRehabilitation*, *5*(1), 3–25.

Fisher, W. P., Jr, Harvey, R. F., Taylor, P., Kilgore, K. M., & Kelly, C. K. (1995b). Rehabits: A common language of functional assessment. *Archives of Physical Medicine and Rehabilitation*, *76*(2), 113–122.

Fisher, W. P., Jr, Melin, J., & Möller, C. (2021). *Metrology for climate-neutral cities* (RISE Research Institutes of Sweden AB No. *RISE Report* 2021:84). Gothenburg, Sweden: RISE. http://ri.diva-portal.org/smash/record.jsf?pid=diva2%3A1616048&dswid=-7140 (79 pp.)

Fisher, W. P., Jr, Oon, E. P.-T., & Benson, S. (2021). Rethinking the role of educational assessment in classroom communities: How can design thinking address the problems of coherence and complexity? *Educational Design Research*, *5*(1), 1–33. doi.org/10.15460/eder.5.1.1537

Fisher, W. P., Jr, Pendrill, L., Lips da Cruz, A., & Felin, A. (2019). Why metrology? Fair dealing and efficient markets for the United Nations' sustainable development goals. *Journal of Physics: Conference Series*, *1379*(012023 http://iopscience.iop.org/article/10.1088/1742-6596/1379/1/012023. doi:10.1088/1742-6596/1379/1/012023

Fisher, W. P., Jr, & Stenner, A. J. (2013). On the potential for improved measurement in the human and social sciences. In Q. Zhang & H. Yang (Eds.). *Pacific Rim Objective Measurement Symposium 2012 Conference Proceedings* (pp. 1–11). Springer-Verlag. https://link.springer.com/chapter/10.1007/978-3-642-37592-7_1

Fisher, W. P., Jr, & Stenner, A. J. (2016). Theory-based metrological traceability in education: A reading measurement network. *Measurement*, *92*, 489–496. http://www.sciencedirect.com/science/article/pii/S0263224116303281

Fisher, W. P., Jr, & Stenner, A. J. (2018a). Ecologizing vs modernizing in measurement and metrology. *Journal of Physics Conference Series, 1044*(012025), http://iopscience.iop.org/article/10.1088/1742-6596/1044/1/012025.

Fisher, W. P., Jr, & Stenner, A. J. (2018b). On the complex geometry of individuality and growth: Cook's 1914 'Curves of Life' and reading measurement. *Journal of Physics Conference Series*, *1065*, 072040 https://iopscience.iop.org/article/10.1088/1742-6596/1065/7/072040/pdf.

Fisher, W. P., Jr, & Wilson, M. (2015). Building a productive trading zone in educational assessment research and practice. *Pensamiento Educativo: Revista de Investigacion Educacional Latinoamericana*, *52*(2), 55–78. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2688260

Fisher, W. P., Jr, & Wilson, M. (2019). The BEAR Assessment System Software as a platform for developing and applying UN SDG metrics. *Journal of Physics Conference Series, 1379*(012041). https://doi.org/10.1088/1742-6596/1379/1/012041

Fisher, W. P., Jr, & Wilson, M. (2020). An online platform for sociocognitive metrology: The BEAR Assessment System Software. *Measurement Science and Technology, 31*(034006). https://iopscience.iop.org/article/10.1088/1361-6501/ab5397/meta

Floyd, J., & Kanamori, A. (2016). Gödel vis-à-vis Russell: Logic and set theory to philosophy. In G. Crocco & E.-M. Engelen (Eds.). *Godelian studies on the Max-Phil notebooks, Vol. I* (pp. 243–326). Presses Universitaires de Provence.

Foxon, T. (2006). Bounded rationality and hierarchical complexity: Two paths from Simon to ecological and evolutionary economics. *Complexity and Ecological Economics*, *3*(4), 361–368.

Fraser, D. A. S. (1963). On sufficiency and the exponential family. *Journal of the Royal Statistical Society: Series B (Methodological)*, *25*(1), 115–123.

French, A. P., & Kennedy, P. J. (1985). *Niels Bohr: A centenary volume*. Harvard University Press.

Frisch, R. (1930). Necessary and sufficient conditions regarding the form of an index number which shall meet certain of Fisher's tests. *Journal of the American Statistical Association*, *25*, 397–406.

Gadamer, H.-G. (1976). *Hegel's dialectic: Five hermeneutical studies*. P. C. Smith (Trans.). Yale University Press.

Gadamer, H.-G. (1979). Historical transformations of reason. In T. F. Geraets (Ed.). *Rationality today* (pp. 3–14). University of Ottawa Press.

Gadamer, H.-G. (1980). *Dialogue and dialectic: Eight hermeneutical studies on Plato*. P. C. Smith, (Trans.). Yale University Press.

Gadamer, H.-G. (1981). *Studies in Contemporary German Social Thought. Vol. 2: Reason in the age of science*. T. McCarthy & F. G. Lawrence (Trans.). MIT Press.

Gadamer, H.-G. (1989). *Truth and method* (Rev. ed.), Crossroad (Original work published 1960).

Gadamer, H.-G. (1991). *Plato's dialectical ethics: Phenomenological interpretations relating to the Philebus* R. M. Wallace (Trans.). Yale University Press.

Galison, P. (1997). *Image and logic: A material culture of microphysics*. University of Chicago Press.

Galison, P. (1999). Trading zone: Coordinating action and belief. In M. Biagioli (Ed.). *The science studies reader* (pp. 137–160). Routledge.

Galison, P. (2008). Image of self. In L. Daston (Ed.). *Things that talk: Object lessons from art and science* (pp. 256–294). Zone Books.

Galison, P., & Stump, D. J. (1996). *The disunity of science: Boundaries, contexts, and power*. Stanford University Press.

Gallaher, M. P., Rowe, B. R., Rogozhin, A. V., Houghton, S. A., Davis, J. L., Lamvik, M. K., & Geikler, J. S. (2007). *Economic impact of measurement in the semiconductor industry* (Tech. Rep. No. 07–2). Gaithersburg, MD: National Institute for Standards and Technology. (191 pages)

Galofaro, F., Toffano, Z., & Doan, B. L. (2018). A quantum-based semiotic model for textual semantics. *Kybernetes*, *47*(2), 307–320.

Gasché, R. (2014). "A certain walk to follow": Derrida and the question of method. *Epoché: A Journal for the History of Philosophy*, *18*(2), 525–550.

Gelven, M. (1984). Eros and projection: Plato and Heidegger. In R. W. Shahan & J. N. Mohanty (Eds.). *Thinking about Being: Aspects of Heidegger's thought* (pp. 125–136). Oklahoma University Press.

George, C., Prigogine, I., & Rosenfeld, L. (1972). *The macroscopic level of quantum mechanics*. Munksgaard.

Gibbs, R. W., Jr. (Ed.). (2008). *Cambridge handbook of metaphor and thought*. Cambridge University Press.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.). *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Lawrence Erlbaum Associates.

Gilligan, C. (1982). *In a different voice*. Harvard University Press.

Glock-Grueneich, N., & Ross, S. N. (2008). Growing the field: The institutional, theoretical, and conceptual maturation of "public participation," part 1: Public participation. *International Journal of Public Participation*, *2*(1), 1–9.

Gnaldi, M., Tomaselli, V., & Forcina, A. (2018). Ecological fallacy and covariates: New insights based on multilevel modelling of individual data. *International Statistical Review*, *86*(1), 119–135.

Golinski, J. (2012). Is it time to forget science? Reflections on singular science and its history. *Osiris*, *27*(1), 19–36.

Graßhoff, U., Holling, H., & Schwabe, R. (2020). D-optimal design for the Rasch counts model with multiple binary predictors. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 541–555.

Green, K. E. (1986). Fundamental measurement: A review and application of additive conjoint measurement in educational testing. *Journal of Experimental Education*, *54*(3), 141–147.

Green, K. E., & Kluever, R. C. (1992). Components of item difficulty of Raven's Matrices. *Journal of General Psychology*, *119*, 189–199.

Grek, S. (2020). Prophets, saviours and saints: Symbolic governance and the rise of a transnational metrological field. *International Review of Education*, *66*(2), 139–166.

Grimby, G., Tennant, A., & Tesio, L. (2012). The use of raw scores from ordinal scales: Time to end malpractice? *Journal of Rehabilitation Medicine*, *44*, 97–98.

Grimm, V., & Railsback, S. F. (2012). Pattern-oriented modelling: A 'multi-scope' for predictive systems ecology. *Philosophical Transactions of the Royal Society B*, *367*, 298–310.

Grosse, M. E., & Wright, B. D. (1986). Setting, evaluating, and maintaining certification standards with the Rasch model. *Evaluation & the Health Professions*, *9*(3), 267–285.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139–150.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.). *Measurement and prediction (Studies in social psychology in World War II. Vol.* 4*)* (pp. 60–90). Wiley.

Guttman, L. (1994). The mathematics in ordinary speech. In S. Levy (Ed.). *Louis Guttman on theory and methodology: Selected writings* (pp. 103–119). Dartmouth Publishing Company.

Habermas, J. (1979). *Communication and the evolution of society*. Beacon Press.

Habermas, J. (1995). *Moral consciousness and communicative action*. MIT Press.

Hambleton, R. K., Swaminathan, H., & Rogers, L. (1991). *Fundamentals of item response theory*. Sage Publications.

Han, Y., Jiang, Z., Ouyang, J., Xu, L., & Cai, T. (2022). Psychometric evaluation of a national exam for clinical undergraduates. *Frontiers in Medicine*, *9*, 1037897.

Hankins, T. L., & Silverman, R. J. (1999). *Instruments and the imagination*. Princeton University Press.

Haraway, D. J. (1996). Modest witness: Feminist diffractions in science studies. In P. Galison & D. J. Stump (Eds.). *The disunity of science: Boundaries, contexts, and power* (pp. 428–441). Stanford University Press.

Harding, S. (1991). *Whose science? Whose knowledge? Thinking from women's lives*. Cornell University Press.

Harding, S. (2008). *Sciences from below: Feminisms, postcolonialities, and modernities*. Duke University Press.

Harding, S., & Hintikka, M. B. (Eds.). (2003). *Discovering reality: Feminist perspectives on epistemology, metaphysics, methodology and philosophy of science, 2nd Ed.* (J. Symons, Ed.). *Studies in epistemology, logic, methodology, and philosophy of science*. Kluwer Academic Publishers.

Hargadon, A. (2003). *How breakthroughs happen: The surprising truth about how companies innovate*. Harvard Business School Press.

Harries, K. (2010). 'Let no one ignorant of geometry enter here': Ontology and mathematics in the thought of Martin Heidegger. *International Journal of Philosophical Studies*, *18*(2), 269–279.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

He, W., Li, S., & Kingsbury, G. G. (2016). A large-scale, long-term study of scale drift: The micro view and the macro view. *Journal of Physics Conference Series*, *772*, 012022. https://iopscience.iop.org/article/10.1088/1742-6596/772/1/012022/meta

Heelan, P. A. (1974). Quantum logic and classical logic: Their respective roles. In R. S. Cohen & M. W. Wartofsky (Eds.). *Logical and epistemological studies in contemporary physics* (pp. 318–349). D. Reidel.

Heelan, P. A. (1983). Natural science as a hermeneutic of instrumentation. *Philosophy of Science*, *50*, 181–204.

Heidegger, M. (1967). *What is a thing?* (Analysis by E. T. Gendlin, Ed.) (W. B. Barton, Jr. & V. Deutsch, Trans.). Regnery/Gateway.

Heidegger, M. (1991). *The principle of reason* R. Lilly (Trans.). Indiana University Press (Original work published 1957).

Heras-Escribano, M. (2020). The evolutionary role of affordances: Ecological psychology, niche construction, and natural selection. *Biology & Philosophy*, *35*, 1–27.

Herrmann-Pillath, C. (2010). Entropy, function and evolution: Naturalizing Peircian semiosis. *Entropy*, *12*(2), 197–242.

Hesse, M. (1970). *Models and analogies in science*. Notre Dame University Press.

Hochstein, A. (2001). A Keynesian view of the Fisher separation theorem. *Atlantic Economic Journal*, *29*(4), 469.

Hodas, N. O., & Lerman, K. (2014). The simple rules of social contagion. *Scientific Reports*, *4*, 4343.

Holton, G. (1988). *Thematic origins of scientific thought: Kepler to Einstein* (rev. ed.), Harvard University Press.

Hotton, S., & Yoshimi, J. (2010). The dynamics of embodied cognition. *International Journal of Bifurcation and Chaos*, *20*(04), 943–972.

Huntley, H. E. (1970). *The divine proportion: A study in mathematical beauty*. Dover.

Huq, A. Z. (2014). The negotiated structural constitution. *Columbia Law Review*, *114*(7), 1595–1686.

Husserl, E. (1970a). *The crisis of European sciences and transcendental phenomenology: An introduction to phenomenological philosophy* D. Carr, (Trans.). Northwestern University Press.

Husserl, E. (1970b). *Philosophie der arithmetik: Mit erganzenden Texten (1890–1901)*. Martinus Nijhoff.

Hutchins, E. (2014). The cultural ecosystem of human cognition. *Philosophical Psychology*, *27*(1), 34–49.

Ifrah, G. (1999). *The universal history of numbers: From prehistory to the invention of the computer*. D. Bellos, I. Monk, E. F. Harding, & S. Wood (Trans.). John Wiley & Sons.

Ihde, D. (1983). The historical and ontological priority of technology over science. In D. Ihde, *Existential technics* (pp. 25–46). State University of New York Press.

Ihde, D. (1991). *Instrumental realism: The interface between philosophy of science and philosophy of technology*. (The Indiana Series in the Philosophy of Technology). Indiana University Press.

Ihde, D. (2002). *Electronic mediations. Vol. 5: Bodies in technology*. University of Minnesota Press.

Irigaray, L. (1984). *An ethics of sexual difference* C. Burke & G. C. Gill, (Trans.). Cornell University Press.

Irigaray, L., & Kuykendall, E. H. (1988). Sorcerer love: A reading of Plato's *Symposium*, Diotima's speech. *Hypatia*, *3*(3), 32–44.

Jaeger, G. (2023). On Wheeler's quantum circuit. In *The quantum-like revolution: A festschrift for Andrei Khrennikov* (pp. 25–59). Springer International Publishing.

Jasanoff, S. (2004). *States of knowledge: The co-production of science and social order*. International Library of Sociology. Routledge.

Jasanoff, S. (2015). Future imperfect: Science, technology, and the imaginations of modernity. In S. Jasanoff & S.-H. Kim (Eds.). *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power* (pp. 1–22). University of Chicago Press.

JCGM: Joint Committee for Guides in Metrology, Working Group 1. (2008). *Guide to the expression of uncertainty in measurement--Evaluation of measurement data*. International Bureau of Weights and Measures--BIPM, Paris.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449–1475.

Karabatsos, G., & Batchelder, W. H. (2003). Markov Chain Monte Carlo estimation theory for test theory without an answer key. *Psychometrika*, *68*(3), 373–389.

Karakus, S., Akpek, E. K., Agrawal, D., & Massof, R. W. (2018). Validation of an objective measure of dry eye severity. *Translational Vision Science & Technology*, *7*(5), 26. https://doi.org/10.1167/tvst.7.5.26

Karasti, H., Millerand, F., Hine, C. M., & Bowker, G. C. (2016). Knowledge infrastructures: Intro to Part I. *Science & Technology Studies*, *29*(1), 2–12.

Kauffmann, S. (1996). *At home in the universe: The search for laws of self-organization and complexity*. Oxford University Press.

Kedarya, T., Elalouf, A., & Cohen, R. S. (2023). Calculating strategic risk in financial institutions. *Global Journal of Flexible Systems Management*, *24*(3), 361–372.

Keller, J., & Tian, P. (2021). The organizational paradox of language. In R. Bednarek, M. E Cunha, J. Schad, & W. K. Smith (Eds.). *Research in the Sociology of Organizations: Vol. 73b. Interdisciplinary Dialogues on Organizational Paradox: Investigating Social Structures and Human Expression, Part B* (pp. 101–122). Emerald Publishing Limited. https://doi.org/10.1108/S0733-558X2021000073b008

Kelley, P. R., & Schumacher, C. F. (1984). The Rasch model: Its use by the National Board of Medical Examiners. *Evaluation & the Health Professions*, *7*(4), 443–454.

Khrennikov, A. (2020). Quantum versus classical entanglement: Eliminating the issue of quantum nonlocality. *Foundations of Physics*, *50*(12), 1762–1780.

Khrennikov, A., & Basieva, I. (2023). Entanglement of observables: Quantum conditional probability approach. *Foundations of Physics*, *53*(5), 84.

King, U. (1989). *The spirit of one earth: Reflections on Teilhard de Chardin and global spirituality*. International Religious Foundation, Incorporated.

Kisiel, T. (2002). The mathematical and the hermeneutical: On Heidegger's notion of the apriori. In A. Denker & M. Heinz (Eds.). *Heidegger's way of thought: Critical and interpretative signposts* (pp. 187–199). Continuum International Publishing Group.

Knorr Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Harvard University Press.

Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, *39*(3), 399–409.

Koopmans, T. C., & Reiersøl, O. (1950). The identification of structural characteristics. *The Annals of Mathematical Statistics*, *XXI*, 165–181.

Kristeva, J. (1980). *Desire in language: A semiotic approach to literature and art*. Columbia University Press.

Kristeva, J. (2014). Reliance, or maternal eroticism. *Journal of the American Psychoanalytic Association*, *62*(1), 69–85.

Kuhn, T. S. (1961). The function of measurement in modern physical science. In T. S. Kuhn, (1977). *The essential tension: Selected studies in scientific tradition and change* (pp. 178–224). University of Chicago Press. (Rpt. from *Isis*, *52(168)*, 161–193. https://www.journals.uchicago.edu/doi/abs/10.1086/349468)

Kuhn, T. S. (1970). *The structure of scientific revolutions*. University of Chicago Press.

Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change*. University of Chicago Press.

Kuhn, T. S. (1993). Metaphor in science. In A. Ortony (Ed.). *Metaphor and thought (2nd Ed.)* (pp. 533–542). Cambridge University Press.

Ladd, H. F. (2017). No child left behind: A deeply flawed federal policy. *Journal of Policy Analysis and Management*, *36*(2), 461–469.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. Basic Books.

Lakoff, G., & Núñez, R. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being*. Basic Books.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard University Press.

Latour, B. (1990). Postmodern? No, simply amodern: Steps towards an anthropology of science. *Studies in History and Philosophy of Science*, *21*(1), 145–171.

Latour, B. (1991). The impact of science studies on political philosophy. *Science, Technology, & Human Values*, *16*(1), 3–19.

Latour, B. (1993). *We have never been modern*. Harvard University Press.

Latour, B. (1996). Not the question. *AAA Anthropology Newsletter*, *37*(3), 1, 5.

Latour, B. (2004). How to talk about the body? The normative dimension of science studies. *Body & Society*, *10*(2–3), 205–229.

Latour, B. (2005). *Reassembling the social: An introduction to Actor-Network-Theory*. Oxford University Press.

Latour, B. (2012). Love your monsters: Why we must care for our technologies as we do our children. *Breakthrough Journal*, *2*, 21–28. Retrieved 1 January 2016, from https://thebreakthrough.org/journal/issue-2/love-your-monsters.

Laszlo, E. (Ed.). (2019). *The new evolutionary paradigm: Keynote volume*. Routledge.

Law, J. (2009). Actor network theory and material semiotics. In B. S. Turner (Ed.). *The new Blackwell companion to social theory* (pp. 141–158). Wiley-Blackwell.

Lenoir, T. (1998). *Inscribing science: Scientific texts and the materiality of communication*. Stanford University Press.

Lerner, R. M., & Overton, W. F. (2017). Reduction to absurdity: Why epigenetics invalidates all models involving genetic reduction. *Human Development*, *60*, 107–123. https://doi.org/10.1159/000477995

Li, G., Pan, Y., & Wang, W. (2021). Using generalizability theory and many-facet Rasch model to evaluate in-basket tests for managerial positions. *Frontiers in Psychology*, *12*, 660553.

Linacre, J. M. (1997). Instantaneous measurement and diagnosis. *Physical Medicine and Rehabilitation State of the Art Reviews, 11*(2), 315–324 [http://www.rasch.org/memo60.htm].

Linssen, R. (1958). *Living Zen*. (D. Abrahams-Curiel, Trans.). George Allen & Unwin. https://terebess.hu/zen/mesterek/Robert-Linssen-Living-Zen.pdf

Lips da Cruz, A., Fisher, W. P. J., Felin, A., & Pendrill, L. (2019). Accelerating the realization of the United Nations Sustainable Development Goals through metrological multi-stakeholder interoperability. *Journal of Physics: Conference Series, 1379*(012046) http://iopscience.iop.org/article/10.1088/1742-6596/1379/1/012046.

Locke, J. (1979). *An essay concerning human understanding, Vols. 1 and 2.* P. H. Nidditch, (Ed.). Oxford University Press.

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review*, *72*(2), 143–155.

Luce, R. D. (1959). *Individual choice behavior*. Wiley.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new kind of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27.

MacArthur, B. D. (2021). Truth and beauty in physics and biology. *Nature Physics*, *17*(2), 149–151.

Mallinson, T. (2024). Extending the justice-oriented, anti-racist framework for validity testing to the application of measurement theory in re(developing) rehabilitation assessments. In W. P. Fisher Jr & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*. De Gruyter.

Mallinson, T., Kozlowski, A. J., Johnston, M. V., Weaver, J., Terhorst, L., Grampurohit, N., Juengst, S., Ehrlich-Jones, L., Heinemann, A. W., Melvin, J., Sood, P., & Van de Winckel, A. (2022). Rasch reporting guideline for rehabilitation research (RULER): The RULER statement. *Archives of Physical Medicine and Rehabilitation*, *103*(7), 1477–1486. https://doi.org/10.1016/j.apmr.2022.03.013

Marais, K., & Kull, K. (2016). Biosemiotics and translation studies. In *Border crossings: Translation studies and other disciplines* (pp. 169–188). Benjamins.

Maran, T. (2007). Towards an integrated methodology of ecosemiotics: The concept of nature-text. *Sign Systems Studies*, *35*(1/2), 269–294.

Maraun, M. D. (1996). Meaning and mythology in the factor analysis model. *Multivariate Behavioral Research*, *31*(4), 603–616.

Mari, L. (2000). Beyond the representational viewpoint: A new formalization of measurement. *Measurement*, *27*, 71–84.

Mari, L. (2003). Epistemology of measurement. *Measurement*, *34*(1), 17–30.

Mari, L. (2021). Is our understanding of measurement evolving? *Acta IMEKO*, *10*(4), 32.

Mari, L., Maul, A., Torres Irribara, D., & Wilson, M. (2016). Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement*, *100*, 115–121. http://www.sciencedirect.com/science/article/pii/S0263224116307497

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement*, *51*, 315–327. http://www.sciencedirect.com/science/article/pii/S0263224114000645

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement, 2nd ed*. (Springer Series in Measurement Science and Technology). Springer. https://link.springer.com/book/10.1007/978-3-031-22448-5

Martin-Lopez, B. (2021). Plural valuation of nature matters for environmental sustainability and justice. *The Royal Society*. Retrieved 5 June 2023, from https://royalsociety.org/topics-policy/projects/biodiversity/plural-valuation-of-nature-matters-for-environmental-sustainability-and-justice/.

Marttila, S. (2016). From rules in use to culture in use: Commoning and infrastructuring practices in an open cultural movement. In P. Lloyd & E. Bohemia (Eds.). *Future Focused Thinking – Design Research Society International Conference 2016, 27 – 30 June* (pp. 454–464). Design Research Society. https://doi.org/10.21606/drs.2016.454

Massof, R. W., & Bradley, C. (2023). An adaptive strategy for measuring patient-reported outcomes. In W. P. Fisher Jr. & S. J. Cano (Eds.). *Person-centered outcome metrology: Principles and applications for high stakes decision making* (pp. 107–150). Springer.

Massof, R. W., Bradley, C., & McCarthy, A. M. (2024). Constructing a continuous latent disease state variable from clinical signs and symptoms, In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*. De Gruyter.

Masters, G. N. (2007). Special issue: Programme for International Student Assessment (PISA). *Journal of Applied Measurement*, *8*(3), 235–335.

Masters, G. N., & Keeves, J. P. (Eds.). (1999). *Advances in measurement in educational research and assessment*. Pergamon.

Maul, A. (2013). On the ontology of psychological attributes. *Theory & Psychology*, *23*(6), 752–769.

Maul, A., Mari, L., & Wilson, M. (2019). Intersubjectivity of measurement across the sciences. *Measurement*, *131*, 764–770.

Maul, A., Torres Irribarra, D., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement*, *79*, 311–320.

McAllister, J. W. (1990). Dirac and the aesthetic evaluation of theories. *Methodology and Science*, *23*(2), 87–102.

McHugh, N., Sinclair, S., Roy, M., Huckfield, L., & Donaldson, C. (2013). Social impact bonds: A wolf in sheep's clothing? *Journal of Poverty and Social Justice*, *21*(3), 247–257.

McLeish, T. (2019). *The poetry and music of science: Comparing creativity in science and art*. Oxford University Press.

McNamara, T. F., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment. Oxford Applied Linguistics)*. Oxford University Press.

Medvedev, O. N., Siegert, R. J., Feng, X. J., Billington, D. R., Jang, J. Y., Krägeloh, & Christian, U. (2016). Measuring trait mindfulness: How to improve the precision of the mindful attention awareness scale using a Rasch model. *Mindfulness*, *7*(2), 384–395.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103–115.

Meijer, D. K. (2015). The universe as a cyclic organized information system: John Wheeler's world revisited. *NeuroQuantology*, *13*(1), 57–78.

Melin, J., Cano, S. J., Flöel, A., Göschel, L., & Pendrill, L. (2023). Metrological advancements in cognitive measurement: A worked example with the NeuroMET memory metric providing more reliability and efficiency. *Measurement: Sensors*, *25*, 100658. https://doi.org/10.1016/j.measen.2022.100658

Meloni, M. (2019). *Impressionable biologies: From the archaeology of plasticity to the sociology of epigenetics*. Routledge.

Meredith, W. M. (1968). *The Poisson distribution and Poisson process in psychometric theory* (Tech. Rep. No. ETS Research Bulletin Series RB-68-42). Princeton, NJ: Educational Testing Service. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1968.tb00565.x (p. 81)

Merleau-Ponty, M. (1964). *The primacy of perception*. J. M. Edie, (Ed.). Northwestern University Press.

Merrell, F. (1996). *Signs grow: Semiosis and life processes*. University of Toronto Press.

Merrell, F. (2011). Tom's often neglected other theoretical source. In P. Cobley, J. Deely, K. Kull, & S. Petrilli (Eds.). *Semiotics continues to astonish: Thomas A. Sebeok and the doctrine of signs* (pp. 251–279). De Gruyter.

Merry, S. E. (2016). *The seductions of quantification: Measuring human rights, gender violence, and sex trafficking*. University of Chicago Press.

Merry, S. E. (2019). The Sustainable Development Goals confront the infrastructure of measurement. *Global Policy*, *10*, 146–148. https://onlinelibrary.wiley.com/doi/pdf/10.1111/1758-5899.12606

Merry, S. E., & Wood, S. (2015). Quantification and the paradox of measurement: Translating children's rights in Tanzania. *Current Anthropology*, *56*(2), 205–229.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, *100*, 398–407.

Miller, P., & O'Leary, T. (2007). Mediating instruments and making markets: Capital budgeting, science and the economy. *Accounting, Organizations, and Society*, *32*(7–8), 701–734.

Mirowski, P. (2004). *The effortless economy of science?* Duke University Press.

Mislevy, R. J. (2014). Postmodern test theory. *Teachers College Record*, *116*(11), 1–24.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A construct-modeling approach to develop a learning progression of how students understand the structure of matter. *Journal of Research in Science Teaching*, *54*(8), 1024–1048.

Morrison, J., & Fisher, W. P., Jr. (2018). Connecting learning opportunities in STEM education: Ecosystem collaborations across schools, museums, libraries, employers, and communities. *Journal of Physics: Conference Series*, *1065*(022009). doi:10.1088/1742-6596/1065/2/022009

Morrison, J., & Fisher, W. P., Jr. (2019). Measuring for management in science, technology, engineering, and mathematics learning ecosystems. *Journal of Physics: Conference Series*, *1379*(012042). doi:10.1088/1742-6596/1379/1/012042

Morrison, J., & Fisher, W. P., Jr. (2021). Caliper: Measuring success in STEM learning ecosystems. *Measurement: Sensors*, *18*, 100327. https://doi.org/10.1016/j.measen.2021.100327

Morrison, J., & Fisher, W. P., Jr. (2024, August 29). *Ecologizing STEM education: Using Caliper to measure and manage for stakeholder empowerment*. Presented at the IMEKO World Congress, Hamburg, Germany.

Moss, P. (2004). The risks of coherence. In M. Wilson (Ed.). *Towards coherence between classroom assessment and accountability* (pp. 217–238). University of Chicago Press.

Mtsatse, N., & Combrinck, C. (2018). Dialects matter: The influence of dialects and code-switching on the literacy and numeracy achievements of isiXhosa Grade 1 learners in the Western Cape. *Journal of Education (University of KwaZulu-Natal)*, *72*, 21–37. https://dx.doi.org/10.17159/2520-9868/i72a02

Mueller, U., & Overton, W. F. (2010). Thinking about thinking – Thinking about measurement: A Rasch analysis of recursive thinking. *Journal of Applied Measurement*, *11*(1), 78–90.

Mueller, U., Sokol, B., & Overton, W. F. (1999). Developmental sequences in class reasoning and propositional reasoning. *Journal of Experimental Child Psychology*, *74*(2), 69–106.

Murphy, M. (2017). *The economization of life*. Duke University Press.

Nagel, E., & Newman, J. R. (1958). *Gödel's proof*. New York University Press.

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, *99*(2), 166–180.

Nersessian, N. J. (1996). Child's play. *Philosophy of Science*, *63*, 542–546.

Nersessian, N. J. (2002). Maxwell and "the method of physical analogy": Model-based reasoning, generic abstraction, and conceptual change. In D. Malament (Ed.). *Reading natural philosophy: Essays in the history and philosophy of science and mathematics* (pp. 129–166). Open Court.

Nersessian, N. J. (2006). Model-based reasoning in distributed cognitive systems. *Philosophy of Science*, *73*, 699–709.

Nersessian, N. J. (2008). *Creating scientific concepts*. MIT Press.

Niederée, R. (1994). There is more to measurement than just measurement: Measurement theory, symmetry, and substantive theorizing. *Journal of Mathematical Psychology*, *38*(4), 527–594.

NIST National Institute for Standards and Technology. (1996). Appendix C: Assessment examples. Economic impacts of research in metrology. In C. o. F. S. Subcommittee on Research (Ed.). *Assessing fundamental science: A report from the Subcommittee on Research, Committee on Fundamental Science*. National Standards and Technology Council. https://wayback.archive-it.org/5902/20150628164643/http://www.nsf.gov/statistics/ostp/assess/nstcafsk.htm#Topic%207

NIST National Institute for Standards and Technology. (2009, December 18). *Outputs and outcomes of NIST laboratory research*. Retrieved 18 April 2020, from NIST: https://www.nist.gov/director/outputs-and-outcomes-nist-laboratory-research.

Noble, D. (2017). Evolution viewed from physics, physiology and medicine. *Interface Focus*, *7*(5), 20160159.

Noddings, N. (1984). *Caring*. University of California Press.

North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System*, *23*(4), 445–465.

North, B. (2014). Putting the Common European Framework of Reference to good use. *Language Teaching*, *47*(2), 228–249. https://doi.org/10.37546/JALTSIG.CEFR2-1

North, B. (2020). Trolls, unicorns and the CEFR: Precision and professionalism in criticism of the CEFR. *CEFR Journal Research and Practice*, *2*, 8–24.

North, D. (1990). A transaction cost theory of politics. *Journal of Theoretical Politics*, *2*(4), 355–367.

Nöth, W. (2018). The semiotics of models. *Sign Systems Studies*, *46*(1), 7–43.

Nöth, W. (2021). System, sign, information, and communication in cybersemiotics, systems theory, and Peirce. In C. Vidales & S. Brier (Eds.). *Introduction to cybersemiotics: A transdisciplinary perspective* (pp. 75–95). Springer.

Nungester, R. J., Dillon, G. F., Swanson, D. B., Orr, N. A., & Powell, R. D. (1991). Standard-setting plans for the NBME comprehensive part I and part II examinations. *Academic Medicine*, *66*(8), 429–433.

Nuzzo, A. (2018). *Approaching Hegel's logic, obliquely: Melville, Moliere, Beckett*. SUNY Press.

Nye, A. (1989). The hidden host: Irigaray and Diotima at Plato's Symposium. *Hypatia*, *3*(3), 45-61. https://www.jstor.org/stable/3809787

Nye, A. (2015). Socrates and Diotima: Sexuality, religion, and the nature of divinity. Springer.

Orr, D. (2006). Diotima, Wittgenstein, and a language for liberation. In D. Orr, L. L. McAlister, E. Kahl & K. Earle (Eds.), *Belief, bodies, and being: Feminist reflections on embodiment* (pp. 59–80). Rowman & Littlefield.

O'Connell, J. (1993). Metrology: The creation of universality by the circulation of particulars. *Social Studies of Science*, *23*, 129–173.

Olteanu, A. (2021). Multimodal modeling: Bridging biosemiotics and social semiotics. *Biosemiotics*, *14*(783–805). https://doi.org/10.1007/s12304-021-09463-7

O'Neill, T., Marks, C. M., & Reynolds, M. (2005). Re-evaluating the NCLEX-RN passing standard. *Journal of Nursing Measurement*, *13*(2), 147–165.

Ostrom, E. (2015). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press (Original work published 1990).

Ottaviani, J., & Purvis, L. (2009). *Suspended in language: Niels Bohr's life, discoveries, and the century he shaped*, 2nd ed*.,* GT Labs.

Overton, W. F. (1994a). The arrow of time and the cycle of time: Concepts of change, cognition, and embodiment. *Psychological Inquiry*, *5*(3), 215–237. https://doi.org/10.1207/s15327965pli0503_9

Overton, W. F. (1994b). Contexts of meaning: The computational and the embodied mind. In W. Overton & D. D. Palermo (Eds.). *The nature and ontogenesis of meaning* (pp. 1–18). Lawrence Erlbaum Associates, Inc.

Overton, W. F. (1997). Beyond dichotomy: An embodied active agent for cultural psychology. *Culture & Psychology*, *3*(3), 315–334.

Overton, W. F. (1998). Developmental psychology: Philosophy, concepts, and methodology. *Handbook of Child Psychology*, *1*, 107–188.

Overton, W. F. (2002). Understanding, explanation, and reductionism: Finding a cure for Cartesian anxiety. In L. Smith & T. Brown (Eds.). *Reductionism and the development of knowledge* (pp. 29–51). Erlbaum.

Overton, W. F. (2007). A coherent metatheory for dynamic systems: Relational organicism-contextualism. *Human Development*, *50*(2/3), 154–159.

Overton, W. F. (2008). Embodiment from a relational perspective. In W. F. Overton, U. Muller, & J. L. Newman (Eds.). *Developmental perspective on embodiment and consciousness* (pp. 1–18). Erlbaum.

Overton, W. F. (2015). Processes, relations and Relational-Developmental-Systems. In W. F. Overton & P. C. M. Molenaar (Eds.). *Theory and Method. Volume 1 of the Handbook of child psychology and developmental science (7th Ed.)* (pp. 9–62). Wiley.

Overwijk, J. (2021). Paradoxes of rationalisation: Openness and control in critical theory and Luhmann's systems theory. *Theory, Culture & Society*, *38*(1), 127–148.

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of Modern Physics*, *87*(3), 925.

Pattee, H. H. (1979). The complementarity principle and the origin of macromolecular information. *Biosystems*, *11*, 217–226.

Pattee, H. H. (2012). Universal principles of measurement and language functions in evolving systems. In H. H. Pattee & J. Raczaszek-Leonardi (Eds.). *Biosemiotics. Vol. 7: Laws, language and life: Howard Pattee's classic papers on the physics of symbols with contemporary commentary* (pp. 181–195). Springer.

Peirce, C. S. (1955). *Philosophical writings of Peirce* (J. Buchler, Ed.). Dover.

Peirce, C. S. (1992). *The essential Peirce: Selected philosophical writings, volume I (1867–1893)*. N. Houser & C. Kloesel, (Eds.). Indiana University Press.

Peirce, C. S., Welby, L. V., Stuart-Wortley, V. A. M. L., & Welby, V. L. (1977). *Semiotic and significs: The correspondence between Charles S. Peirce and Victoria Lady Welby*. C. S. Hardwick & J. Cook, (Eds.). Indiana University Press.

Pendrill, L. R. (2006). Optimised measurement uncertainty and decision-making when sampling by variables or by attribute, *Measurement*, *39*, 829–840, http://dx.doi.org/10.1016/j.measurement.2006.04.014

Pendrill, L. R. (2014). Man as a measurement instrument [Special Feature]. *NCSLi Measure: The Journal of Measurement Science*, *9*(4), 22–33. http://www.tandfonline.com/doi/abs/10.1080/19315775.2014.11721702

Pendrill, L. R. (2019). *Quality assured measurement: Unification across social and physical sciences*. Springer Series in Measurement Science and Technology). Springer. https://link.springer.com/book/10.1007/978-3-030-28695-8

Pendrill, L., & Fisher, W. P., Jr. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement*, *71*, 46–55. http://dx.doi.org/10.1016/j.measurement.2015.04.010

Pendrill, L. R., & Melin, J. (2023). Assuring measurement quality in person-centred care. In W. P. Fisher Jr. & S. J. Cano (Eds.). *Person-centered outcome metrology* (pp. 311–355). Springer.

Pennecchi, F. R., Kuselman, I., Hibbert, D. B., Sega, M., Rolle, F., & Altshul, V. (2022). Fit-for-purpose risks in conformity assessment of a substance or material–A case study of synthetic air. *Measurement*, *188*, 110542.

Pepper, S. C. (1942). *World hypotheses: A study in evidence*. University of California Press.

Petersen, A. (1968). *Quantum physics and the philosophical tradition*. MIT Press.

Petracca, E., & Gallagher, S. (2020). Economic cognitive institutions. *Journal of Institutional Economics*, *16*(6), 747–765.

Petrosky, T., & Prigogine, I. (1990). Laws and events: The dynamical basis of self-organization. *Canadian Journal of Physics*, *68*(9), 670–682.

Platt, J. R. (1961). Social chain reactions. *Bulletin of the Atomic Scientists: Man and His Habitat, Part II*, *17*, 365–386. (Rpt. in J. R. Platt, 1961, *The step to man* (pp. 39–52). John Wiley & Sons.) http://dx.doi.org/10.1080/00963402.1961.11454270

Polanyi, M. (1974). *Personal knowledge: Towards a post-critical philosophy*. University of Chicago Press.

Pollitt, J. J. (1972). *Art and experience in classical Greece*. Cambridge University Press.

Popescu, C. (2016). From laboratory life to the living and tinkering laboratories of care: A new perspective in STS research? *EASST Review*, *35*(4). https://easst.net/issue/easst-review-volume-354-december-2016/

Poposki, N., Majcen, N., & Taylor, P. (2009). Assessing publically financed metrology expenditure against economic parameters. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement*, *14*(7), 359–368.

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Porter, T. M. (1999). Quantification and the accounting ideal in science. In M. Biaglioli (Ed.). *The science studies reader* (pp. 394–406). Routledge.

Postman, N. (1992). *Technopoly: The surrender of culture to technology*. Vintage Books.

Power, M. (2004). Counting, control, and calculation: Reflections on measuring and management. *Human Relations*, *57*(6), 765–783. doi:10.1177/0018726704044955

Powers, M., & Fisher, W. P., Jr. (2021). Physical and psychological measures quantifying functional binocular vision. *Measurement: Sensors*, *18*, 100320. https://doi.org/10.1016/j.measen.2021.100320

Prigogine, I. (1976). Order through fluctuation: Self-organization and social system. In E. Jantsch & C. Waddington (Eds.). *Consciousness and evolution: Human systems in transition* (pp. 93–130). Addison Wesley.

Prigogine, I. (1986). Science, civilization and democracy: Values, systems, structures and affinities. *Futures*, *18*(4), 493–507.

Prigogine, I. (1997). *The end of certainty: Time, chaos, and the new laws of nature*. Free Press.

Prigogine, I., & Allen, P. M. (1982). The challenge of complexity. In W. C. Schieve & P. Allen (Eds.). *Self-organization and dissipative structures: Applications in the physical and social sciences* (pp. 3–39). University of Texas Press.

Prigogine, I., & Lefever, R. (1973). Theory of dissipative structures. In H. Haken (Ed.). *Synergetics: Cooperative phenomena in multi-component systems* (pp. 124–135). Springer.

Prigogine, I., & Stengers, I. (1984). *Order out of chaos: Man's new dialogue with nature*. Bantam Books.

Puig de la Bellacasa, M. (2015). Ecological thinking, material spirituality, and the poetics of infrastructure. In G. C. Bowker, S. Timmermans, A. E. Clarke, & E. Balka (Eds.). *Boundary objects and beyond: Working with Leigh Star* (pp. 47–68). MIT Press.

Pugliese, K., Fisher, W. P., Jr, Kelly, C. K., Accardi, R., & Harvey, R. F. (1993). Accountability in pastoral care: Can spiritual well-being be measured? [abstract]. *Archives of Physical Medicine and Rehabilitation*, *74*, 1270.

Quaglia, M., Pendrill, L., Melin, J., & Cano, S., & 15HLT04 NeuroMET Consortium. (2016–2019). *Innovative measurements for improved diagnosis and management of neurodegenerative diseases* (EMPIR NeuroMET). Teddington, Middlesex, UK: Euramet. https://www.lgcgroup.com/our-programmes/empir-neuromet/neuromet-landing-page/ (p. 36)

Quaglia, M., Pendrill, L., Melin, J., & Cano, S., & 18HLT09 NeuroMET2 Consortium. (2019–2022). *Publishable Summary for 18HLT09 NeuroMET2: Metrology and innovation for early diagnosis and accurate stratification of patients with neurodegenerative diseases* (EMPIR NeuroMET). Teddington, Middlesex, UK: EURAMET. https://www.lgcgroup.com/our-programmes/empir-neuromet/neuromet-landing-page/ (p. 5).

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Danmarks Paedogogiske Institut.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.). *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability: Volume IV:*

Contributions to biology and problems of medicine (pp. 321–333. University of California Press, http://www.rasch.org/memo1960.pdf

Rasch, G. (1968). *A mathematical theory of objectivity and its consequences for model construction.* Unpublished paper [http://www.rasch.org/memo1968.pdf] presented at the European Meeting on Statistics, Econometrics and Management Science. Institute of Mathematical Statistics, European Branch, Amsterdam, the Netherlands, September 6.

Rasch, G. (1972/1988). Review of the cooperation of Professor B. D. Wright, University of Chicago, and Professor G. Rasch, University of Copenhagen; letter of June 18, 1972. *Rasch Measurement Transactions*, *2*(2), 19 http://www.rasch.org/rmt/rmt22c.htm.

Rasch, G. (1972/2010). Retirement lecture of 9 March 1972: Objectivity in social sciences: A method problem. (Cecilie Kreiner, Trans.). *Rasch Measurement Transactions*, *24*(1), 1252–1272. http://www.rasch.org/rmt/rmt241.pdf.

Rasch, G. (1973/2011). All statistical models are wrong! Comments on a paper presented by Per Martin-Löf, at the Conference on Foundational Questions in statistical inference, Aarhus, Denmark, May 7–12, 1973. *Rasch Measurement Transactions*, *24*(4), 1309. http://www.rasch.org/rmt/rmt244.pdf.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, *14*, 58–94. https://www.rasch.org/memo18.htm

Richards, R. J., & Daston, L. (Eds.). (2019). *Kuhn's' structure of scientific revolutions' at fifty: Reflections on a science classic*. University of Chicago Press.

Ricoeur, P. (1965). *History and truth* C. A. Kelbley, (Trans.). Northwestern University Press.

Ricoeur, P. (1974a). The project of a social ethic. In D. Stewart & J. Bien, (Eds.). *Political and social essays by Paul Ricoeur* (pp. 160–175). Ohio University Press.

Ricoeur, P. (1974b). Violence and language. In D. Stewart & J. Bien (Eds.). *Political and social essays by Paul Ricoeur* (pp. 88–101). Ohio University Press.

Ricoeur, P. (1977). *The rule of metaphor: Multi-disciplinary studies of the creation of meaning in language* R. Czerny, (Trans.). University of Toronto Press.

Ricoeur, P. (1981). The model of the text: Meaningful action considered as a text. In J. B. Thompson (Ed.). *Hermeneutics and the human sciences: Essays on language, action and interpretation* (pp. 197–221). Cambridge University Press.

Ricoeur, P. (2020). *Philosophy, ethics, and politics*. John Wiley & Sons.

Rogosa, D. (1987). Casual [sic] models do not support scientific conclusions: A comment in support of Freedman. *Journal of Educational Statistics*, *12*(2), 185–195.

Roosevelt, E. (1960/2011). *You learn by living*. Harper Perennial.

Ross, S. N. (2007). The case for developmental methodologies in democratization. *Journal of Adult Development*, *14*(3–4), 80–90. doi 10.1007/s10804-007-9015-6

Ross, S. N. (2008). A future society functioning at the paradigmatic stage? *World Futures*, *64*(5–7), 554–562.

Ross, S. N. (2014). A developmental behavioral analysis of dual motives' role in political economies of corruption. *Integral Review: A Transdisciplinary & Transcultural Journal for New Thought, Research, & Praxis*, *10*(1), 91–121.

Ross, S. N., & Commons, M. L. (2008). Applying hierarchical complexity to political development. *World Futures*, *64*, 480–497.

Ross, S. N., & Glock-Grueneich, N. (2008a). Growing the field: The institutional, theoretical, and conceptual maturation of "public participation." Part 3: Theoretical maturation. *International Journal of Public Participation*, *2*(1), 14–25.

Ross, S. N., & Glock-Grueneich, N. (2008b). Growing the field: The institutional, theoretical, and conceptual maturation of "public participation," part 5: Implications for IJP2 serving the field. *International Journal of Public Participation*, *2*(1), 30–32.

Rousseau, D. M. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. *Research in Organizational Behavior*, *7*(1), 1–37.

Roubach, M. (2008). *Being and number in Heidegger's thought*. Continuum.

Ruddick, S. (1989). *Maternal thinking*. Beacon Press.

Rudrauf, D., Bennequin, D., Granic, I., Landini, G., Friston, K., & Williford, K. (2017). A mathematical model of embodied consciousness. *Journal of Theoretical Biology*, *428*, 106–131.

Russell, B. (1948). *Human knowledge: Its scope and limits*. Allen & Unwin.

Russell, M. (2023). Shifting educational measurement from an agent of systemic racism to an anti-racist endeavor. *Applied Measurement in Education [Special Issue on Advancing Racial Justice in Educational Assessment: Perspectives and Practices]*, *36*(3), 216–241. https://doi.org/10.1080/08957347.2023.2217555

Sahlins, M. (2022). *The new science of the enchanted universe: An anthropology of most of humanity*. Princeton University Press.

Salsburg, D. S. (1985). The religion of statistics as practiced in medical journals. *The American Statistician*, *39*(3), 220–223.

San Martin, E., Gonzalez, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3 PL model. *Psychometrika*, *80*(2), 450–467.

San Martin, E., Perticará, M., Varas, I. M., Kenzo Asahi, K., & González, J. (2024). The role of identifiability in empirical research. In W. P. Fisher Jr. & L. R. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*. De Gruyter.

San Martin, E., & Rolin, J. M. (2013). Identification of parametric Rasch-type models. *Journal of Statistical Planning and Inference*, *143*(1), 116–130.

Schaffer, S. (1992). Late Victorian metrology and its instrumentation: A manufactory of Ohms. In R. Bud & S. E. Cozzens (Eds.). *Invisible connections: Instruments, institutions, and science* (pp. 23–56). SPIE Optical Engineering Press.

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press.

Sebeok, T. A. (1991). *A sign is just a sign*. Indiana University Press.

Sebeok, T. A. (2001). *Signs: An introduction to semiotics*. University of Toronto Press.

Seiler, E. (1999). The role of metrology in economic and social development. *Computer Standards & Interfaces*, *21*, 77–88.

Semerjian, H. G., & Watters, R. L. (2000). Impact of measurement and standards infrastructure on the national economy and international trade. *Measurement*, *27*(3), 179–196.

Sendak, M. (1963). *Where the Wild Things Are*. Harper & Row.

Shanahan, M. (2014). A pregnant space: Levinas, ethics and maternity. In M. Shanahan (Ed.). *An ethics for/ of the future?* (pp. 116–126). Cambridge Scholars Publishing.

Shapin, S. (1989). The invisible technician. *American Scientist*, *77*(6), 554–563.

Shapin, S., & Schaffer, S. (1985). *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life*. Princeton University Press.

Sijtsma, K. (2016). Playing with data – or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika*, *81*(1), 1–15.

Sinclair, S., McHugh, N., & Roy, M. J. (2019). Social innovation, financialisation and commodification: A critique of social impact bonds. *Journal of Economic Policy Reform*, *24*(1), 11–27.

Smith, J. E., & Nau, R. F. (1995). Valuing risky projects: Option pricing theory and decision analysis. *Management Science*, *41*(5), 795–816.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational & Psychological Measurement*, *51*, 541–565.

Smith, R. M. (1996a). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling*, *3*(1), 25–40.

Smith, R. M. (1996b). Item component equating. In M. Wilson (Ed.). *Objective measurement: Theory into practice, Vol.* 3 (pp. 289–308). Ablex Publishing Co.

Smith, R. M., Julian, E., Lunz, M., Stahl, J., Schulz, M., & Wright, B. D. (1994). Applications of conjoint measurement in admission and professional certification programs. *International Journal of Educational Research*, *21*(6), 653–664.

Smolin, L. (2014). *Time reborn: From the crisis of physics to the future of the universe*. Allan Lane.

Snowden, A., Karimi, L., & Tan, H. (2022). Statistical fit is like beauty: A Rasch and factor analysis of the Scottish PROM. *Journal of Health Care Chaplaincy*, *28*(3), 415–430.

Solloway, S., & Fisher, W. P., Jr. (2007). Mindfulness in measurement: Reconsidering the measurable in mindfulness. *International Journal of Transpersonal Studies*, *26*, 58–81 http://digitalcommons.ciis.edu/ijts-transpersonalstudies/vol26/iss1/8.

Star, S. L. (1988). The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving. *Proceedings of the 8th AAAI Workshop on Distributed Artificial Intelligence, Technical Report, Department of Computer Science, University of Southern California.* (Rpt. in G. Bowker, S. Timmermans, A. E. Clarke & E. Balka, (Eds.). (2015). *Boundary objects and beyond: Working with Leigh Star* (pp. 243–259). The MIT Press).

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, ‘translations,’ and boundary objects: Amateurs and professionals in Berkeley’s Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science*, *19*(3), 387–420. (Rpt. in G. Bowker, S. Timmermans, A. E. Clarke & E. Balka, (Eds.). (2015). Boundary objects and beyond: Working with Leigh Star (pp. 171–200). The MIT Press).

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, *7*(1), 111–134. (Rpt. in G. Bowker, S. Timmermans, A. E. Clarke & E. Balka, (Eds.). (2015). Boundary objects and beyond: Working with Leigh Star (pp. 377–415). The MIT Press)

Stenner, A. J., Fisher, W. P., Jr, Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology: Quantitative Psychology and Measurement*, 4(536), 1–14. doi: 10.3389/fpsyg.2013.00536 (Rpt. in Fisher, W. P. Jr., & Massengill, P. J. *Explanatory models . . .* (pp. 223–250). Springer, 2023).

Stenner, A. J., & Stone, M. (2010). Generally objective measurement of human temperature and reading ability: Some corollaries. *Journal of Applied Measurement*, 11(3), 244–252. (Rpt. in W. P. Fisher, Jr. & P. J. Massengill (Eds.). *Explanatory models . . .* (pp. 166–177). Springer, 2023).

Stiglitz, J. E., Sen, A., & Fitoussi, J.-P. (2010). *Mismeasuring our lives: Why GDP doesn’t add up*. The New Press.

Sul, D. (2024). Situating culturally specific assessment development within the disjuncture-response dialectic. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI* (p. in press). De Gruyter.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.). *Handbook of mathematical psychology* (pp. 1–76). Wiley.

Susser, M., & Susser, E. (1996). Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemiology [see comments] [published erratum appears in Am J Public Health 1996 Aug;86(8 Pt 1):1093]. *American Journal of Public Health*, *86*(5), 674–677.

Sutton, J., Harris, C. B., Keil, P. G., & Barnier, A. J. (2010). The psychology of memory, extended cognition, and socially distributed remembering. *Phenomenology and the Cognitive Sciences*, *9*(4), 521–560.

Swimme, B. T., & Tucker, M. E. (2011). *Journey of the universe: An epic story of cosmic, earth, and human transformation*. Yale University Press.

Theriault, J. E., Young, L., & Barrett, L. F. (2021). The sense of should: A biologically-based framework for modeling social pressure. *Physics of Life Reviews*, *36*, 100–136.

Thurstone, L. L. (1926). The scoring of individual performance. *Journal of Educational Psychology*, *17*, 446–457.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, XXXIII, 529–544. (Rpt. in L. L. Thurstone, *The measurement of values* (pp. 215–233). University of Chicago Press, Midway Reprint Series, 1959).

Thurstone, L. L. (1937). Psychology as a quantitative rational science. *Science*, *LXXXV (*March 5*)*, 228–232. (Rpt., L. L. Thurstone, *The Measurement of Values*, Midway Reprint Series. University of Chicago Press, 1959, pp. 3–11).

Toulmin, S. E. (1953). *The philosophy of science: An introduction*. Hutchinson's University Library.

Toulmin, S. E. (1961). *Foresight and understanding: An enquiry into the aims of science*. Hutchinson.

Toulmin, S. E. (1982). The construal of reality: Criticism in modern and postmodern science. *Critical Inquiry*, *9*, 93–111.

Townes, C. H. (2001). Logic and uncertainties in science and religion. In *Science and the future of mankind: Science for man and man for science* (pp. 296–309). Pontificia Academia Scientiarum.

Townes, C. H., Merritt, F. R., & Wright, B. D. (1948). The pure rotational spectrum of ICL. *Physical Review*, *73*, 1334–1337.

Tozzi, A. (2021). An approach to pluralizing socionatural resilience through assemblages. *Progress in Human Geography*, *45*(5), 1083–1104.

Tsai, T.-H., Kramer, G. A., Yang, C.-L., Neumann, L. M., & Chang, S.-R. (2013). NBDE Part II practice analyses: An overview. *Journal of Dental Education*, *77*(12), 1566–1580.

Tymieniecka, A.-T. (1998). The ontopoiesis of life as a new philosophical paradigm. *Phenomenological Inquiry*, *22*, 12–59.

Ulbrich, P., Porto de Albuquerque, J., & Coaffee, J. (2019). The impact of urban inequalities on monitoring progress towards the sustainable development goals: Methodological considerations. *ISPRS International Journal of Geo-Information*, *8*(1), 6.

van Fraassen, B. C. (1974). The labyrinth of quantum logics. In R. S. Cohen & M. W. Wartofsky (Eds.). *Logical and epistemological studies in contemporary physics* (pp. 224–254). D. Reidel.

van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford University Press.

von Neumann, J. (1955). *Mathematical foundations of quantum mechanics*. Princeton University Press.

Vygotsky, L. S. (1978). *Mind and society: The development of higher mental processes*. Harvard University Press.

Wagner, A. (2023). *Sleeping beauties: The mystery of dormant innovations in nature and culture*. Simon and Schuster.

Watts, A. (1973). *Lectures on nature of man as a unified organism-environment field*. Sausalito, CA: The Society for Comparative Philosophy, Inc.

Weinsheimer, J. (1985). *Gadamer's hermeneutics: A reading of Truth and Method*. Yale University Press.

Weiss, D. J. (2021). *The Rasch model*. Retrieved 6 March 2024, from Assessment System Corp.: https://assess.com/the-rasch-model/.

Weitzel, T. (2004). *Economics of standards in information networks*. Physica-Verlag.

Weitzel, M. J., & Johnson, W. M. (2012). Using target measurement uncertainty to determine fitness for purpose. *Accreditation and Quality Assurance*, *17*(5), 491–495.

Wendt, A., & Tatum, D. S. (2005). Credentialing health care professionals. In N. Bezruczko (Ed.). *Rasch measurement in health sciences* (pp. 161–175). JAM Press.

Wheeler, J. A. (1974). The universe as a home for man. *American Scientist*, *62*, 683–691. https://www.jstor.org/stable/27845173

Wheeler, J. A. (1980). Law without law. In P. Medawar & J. Shelley (Eds.). *Structure in science and art* (pp. 132–154). Excerpta Medica.

Wheeler, J. A. (1981). The participatory universe. *Science*, *81*(2), 5.

Wheeler, J. A. (1982). Particles and geometry. In Unified theories of elementary particles: Critical assessments and prospects [Special issue]. *Lecture Notes in Physics*, *160*, 189–217.

Wheeler, J. A. (1988). World as system self-synthesized by quantum networking. *IBM Journal of Research and Development*, *32*(1), 4–15.

Wheeler, J. A. (1994). *At home in the universe*. American Institute of Physics Press.

Wheeler, J. A. (2014). Law without law. In J. A. Wheeler & W. Zurek (Eds.). *Quantum theory and measurement* (pp. 182–213). Princeton University Press. (Originally published in 1983).

Wheeler, J. A. (2018). Information, physics, quantum: The search for links. In *Feynman and computation: Exploring the limits of computers* (pp. 309–336). CRC Press.

Wheeler, J. A., & Zurek, W. H. (Eds.). (1983/2014). *Quantum theory and measurement*. Princeton University Press.

Whitehead, A. N. (1911). *An introduction to mathematics*. Henry Holt and Co.

Whitehead, A. N. (1925). *Science and the modern world*. Macmillan.

Wigner, E. P. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Communications on Pure and Applied Mathematics*, *13*, 1–14.

Wilson, M. R. (Ed.). (2004). *National Society for the Study of Education Yearbooks. Vol. 103, Part II: Towards coherence between classroom assessment and accountability.* University of Chicago Press.

Wilson, M. R. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Wilson, M. R. (2013a). Seeking a balance between the statistical and scientific elements in psychometrics. *Psychometrika*, *78*(2), 211–236.

Wilson, M. R. (2013b). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement*, *46*, 3766–3774. http://www.sciencedirect.com/science/article/pii/S0263224113001061

Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice*, *37*(1), 5–20.

Wittgenstein, L. (1922). *Tractatus logico-philosophicus*. Harcourt Brace.

Wittgenstein, L. (1958). *Philosophical investigations* G. E. M. Anscombe, (Trans.); 3d ed.), Macmillan.

Wittgenstein, L. (1983). *Remarks on the foundations of mathematics*. G. H. von Wright, R. Rhees, G. E. M. Anscombe, & G. E. M. Anscombe, (Trans.). MIT Press.

Wright, B. D. (1958). On behalf of a personal approach to learning. *The Elementary School Journal*, *58*(7), 365–375.

Wright, B. D. (1965). Estimating Rasch models for measurement. In J. Loevinger (Chair), *Symposium on Sample Free Probability Models for Psychosocial Measurement*. Midwestern Psychological Association, Chicago, IL, April.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 invitational conference on testing problems* (pp. 85–101). http://www.rasch.org/memo1.htm. Educational Testing Service.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97–116. http://www.rasch.org/memo42.htm.

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review*, *3*(1), 281–288. http://www.rasch.org/memo41.htm.

Wright, B. D. (1994). Measuring and counting. *Rasch Measurement Transactions*, *8*(3), 371. http://www.rasch.org/rmt/rmt83c.htm.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*(4), 33–45, 52. https://doi.org/10.1111/j.1745-3992.1997.tb00606.x

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. MESA Press.

Wright, B. D., Mead, R. J., & Ludlow, L. H. (1980). *KIDMAP: person-by-item interaction mapping* (Tech. Rep. No. MESA Memorandum #29). Chicago: MESA Press [http://www.rasch.org/memo29.pdf]. (p. 6)

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. MESA Press.

Zhang, X., & Roberts, W. L. (2012). Investigation of standardized patient ratings of humanistic competence on a medical licensure examination using Many-Facet Rasch Measurement and generalizability theory. *Advances in Health Sciences Education*, *18*(5), 929–944. https://doi.org/10.1007/s10459-012-9433-5

Zukav, G. (1979). *The dancing Wu-Li masters: An overview of the new physics*. Bantam Books.

Part II: **Designing and Calibrating Metrologically Viable Measurements: Methods and Applications**

Robert W. Massof, Chris Bradley, and Allison M. McCarthy

7 Constructing a continuous latent disease state variable from clinical signs and symptoms

Abstract: The concept of a syndrome, disorder, or disease implies the identification of a particular etiological or pathophysiological process defined by an ordered collection of signs and symptoms, each of which begins at its own time and progresses in severity at its own rate. The signs and symptoms per se are disease indicator variables that can be measured in physical units or scaled subjectively with ordered categories. This chapter shows how the unique pattern and dynamics of the progression of different disease indicators can be used to estimate the magnitude of a single latent disease state variable for a patient. In so doing, the methods used here exemplify the application of metrological measurement modeling of an interval scale for a complex disease state combining physical measurements and ordered categorical scores. The enabling assumption is that despite the diversity of indicator variables observed and recorded in different units, the single latent disease state variable is a monotonic multivariable function of a vector of indicator variables. We report an approach that scales relative frequency distributions of observed indicator variables, specified in their unique units, for a sample of patients to equal entropy units, which equates uncertainty for each observation. An axiomatic polytomous probabilistic conjoint measurement model is then employed to estimate the patient's latent disease state from the entropy in the observed vector of disease signs and symptoms. An example of the analysis is illustrated with clinical observations of dry eye disease signs and symptoms. The results of the analysis show that the invariance of measurements constructed in this way exhibits the kinds of properties necessary and sufficient to quality-assured unit standards.

Keywords: conjoint measurementl, disease progression, dry eye, entropy, Rasch model

7.1 Introduction

Patients typically present to the healthcare system with self-identified signs and symptoms of their malady. The physician elaborates on the patient's complaints by eliciting a

Robert W. Massof, Department of Ophthalmology, Department of Neuroscience, School of Medicine, Johns Hopkins University, Baltimore, Maryland, U.S.A.

Chris Bradley, Department of Ophthalmology, School of Medicine, Johns Hopkins University, Baltimore, Maryland, U.S.A.

Allison M. McCarthy, Department of Psychiatry and Behavioral Sciences, School of Medicine, Vanderbilt University and Center for Biomedical Ethics and Society, Vanderbilt University Medical Center, Nashville, Tennessee, U.S.A.

https://doi.org/10.1515/9783111036496-007

structured history, performing a physical examination, obtaining tests and measurements in the form of quantitative data about anatomical, physiological, and psychological status, and employing their pattern recognition skills and/or some other personal hypothesis testing strategy to diagnose the health problem and develop a plan of care. The pattern recognized by the physician in the patient's data constitutes a clinical *syndrome*. If a syndrome can be narrowed to an abnormal anatomical structure or physiological process, the term *disorder* is used. If the cause of the disorder can be identified, then the clinical syndrome takes on the status of a *disease*. One unifying feature of these conceptual categories is a theoretical and practical commitment to perceiving the constellation of signs and symptoms as progressing along a relatively predictable course, unless intervention occurs; this commitment is central to the clinical activity of *diagnosis*. Interventions can reverse signs and/or reduce symptoms (palliative treatments) and/or they can serve to cure the disease (curative treatments), or at least slow or otherwise alter its progression (Stegenga, 2018). To the extent that a disorder or disease follows a natural statistical course, it can be characterized as an evolution of pathognomonic signs and symptoms.

We propose that the ordering of the appearance of different signs and symptoms and of their changes as the disease progresses can be used to construct an objective metric of a putative latent disease state variable on a continuous equal interval scale (in this chapter, the distinction between disorder and disease is inconsequential with respect to our goals, that is, disorders are regarded as diseases with unidentified causes, so going forward we take the liberty of using these terms interchangeably). The aims of this chapter are to develop the concept, illustrate examples of applications, and demonstrate a method of estimating objective measurements of latent disease state variables from diverse clinical observations. We draw our examples from ophthalmology, the clinical discipline in which the authors have access to and experience with appropriate diagnostic data.

7.2 Clinical observations of manifest disease state variables

As used in this chapter, the term *variable* is defined to be a scalable property or trait subject to measurement, more specifically called the *measurand,* and as the mathematical representation of the quantity assigned to the measurand (denoted with a number or a symbol that represents a number). The assignment of a quantity to the measurand requires a *measurement instrument* (i.e., a set of operations that equates observations of the measurand with a sum of standard units that represents the quantity or magnitude of the observed property or trait) (Pendrill, 2019). In the physical sciences, two types of variables are identified: *extensive*, such as length, which can be observed and measured directly as a count of concatenated or summed units of equal magnitude that match the observation; and *intensive*, which cannot be observed directly and compared to a count of equal physical units. Intensive variables, such as temperature, are inferred from observations of their effects on extensive variables, and are constructed from, and measured by way of theory (even an obsolete theory such as the caloric theory of heat in the case of temperature). Measurements of intensive variables consist of operations that correspond to mathematical relationships between the estimated variable and measurements of observed effects on extensive variables (e.g., measuring temperature by measuring its effects on the length of a column of mercury in a closed capillary tube).

Analogously, in the behavioral and social sciences, variables are classified as manifest or latent. Manifest variables include any physical variable, both extensive and intensive, that can be observed and agreed upon publicly. Manifest variables also can include the publicly observable reports per se of subjective magnitude estimates made by individual patients (such as a pain score), or by second party observers (such as an Apgar score). However, unlike intensive variables in physics, at best the reported magnitude estimates must be regarded by the public as private, unverifiable, ordering of observations made by a judge. In this case, the value reported by the judge is manifest, but the variable being judged is latent because it cannot be observed directly and compared to a count of publicly observable concatenated standard units. Like intensive variables, the magnitude of latent variables must be inferred by way of theory from observations of manifest variables that can be ordered, which in our case include patient reports and ordinal clinician judgments (Massof, 2002).

Signs and symptoms are manifest variables that can be ordered by magnitude (e.g., physical dimensions, strength, or count). When observed for the purpose of estimating measurements of latent variables, manifest variables, even those thought of as being continuous, must be discretized. If the unit divisions defining the scale are ordered, concatenated, equal in magnitude, and additive, then the scale is considered *equal interval*, which can be represented with an elementary number line (Michell, 1990). Strictly speaking, measurements ultimately require the scales to be expressed as counts of equal units that can be added to match the magnitude of the measurand for the property or trait of interest.

Diagnosing a patient involves inference of the presence of an underlying disease. This inference is based first and foremost on observation of signs and symptoms manifest in the patient (consistent with the notion that diagnosis is a central example of "inference to the best explanation"). The attribution of a disease to a patient, however, must be able to survive change in the patient's clinical presentation; otherwise, the concept of disease is nothing more than the naming of a specific constellation of signs and symptoms at an observed time. This resonates with a common intellectual historical account of the modern practice of disease categorization as growing out of depersonalized observations of consistent patterns in signs and symptoms and their progression in severity in groups of patients with similar case histories who were being cared for in teaching clinics and hospital wards (i.e., a product of Foucault's "medical gaze") (Foucault, 1973). We can refer to the successful survival of a diagnosis despite substantial change in the patient's clinical presentation as "progression" of the disease.

"Disease state" is a hypothetical continuous latent variable that corresponds to a magnitude estimate by the clinician of the disease progression (Massof & McDonnell, 2012). Whether referring to the four ordinal grades of cancer or to the four ordinal grades of cataract, the signs and symptoms of a disease are manifest variables and are assumed to progress through ordered stages. Different signs and symptoms may become detectable at different times over the natural progression of disease severity and, once detected, may progress at different rates. Theoretically, the different signs and symptoms can be considered elements of a time-dependent disease state vector. Mathematically, we think of disease progression as represented by monotonic functions for the worsening of signs and symptoms over time, but there could be remissions in selected signs or symptoms that would result in nonmonotonicities in the mathematical description of their natural course. Recovery from disease signs and symptoms, whether natural or in response to treatment, may follow a different pattern, rather than simply being the reverse of the dynamic pattern seen in the natural progression of the disease. This caveat is particularly applicable to palliative treatments.

The concept of employing an array of ordered clinical observations to estimate a latent disease state variable on a continuous equal interval scale is most relevant to clinical research – disease natural history studies, controlled treatment trials, and comparative effectiveness studies. Current practice is to employ a single primary outcome measurement (e.g., visual acuity) or a dichotomous endpoint (e.g., development of retinal neovascularization) to draw conclusions about the efficacy of an intervention. Even in cases that employ a vector of parameterized measurements, for example, perimetric measurements of visual sensitivity (i.e., samples of visual sensitivity at different locations throughout the visual field), the multiple measurements (vector elements) are reduced to a single summary variable (e.g., mean deviation from age-matched "normal" values in the case of perimetric visual sensitivity measurements) for the purpose of creating a single primary outcome or an endpoint variable. However, most prospective clinical research studies, especially those conducted on novel treatments to obtain approval from regulatory agencies (e.g., US Food and Drug Administration, European Medicines Agency), include many more observations than simply the one identified as the primary outcome or endpoint. These "secondary" observations are typically used to corroborate conclusions from, or to frame the discussion of, the primary observation.

We tend to think of the battery of clinical observations, mediated by the background attribution of a specific disease to a patient, as indicators of the patient's disease state. Some variables may be indicators of early changes in the disease state, whereas other variables may remain normal over much of the course of the disease and then exhibit changes in value at later stages. For example, a group of pigmentary retinal degenerations (presumed to be inherited) called retinitis pigmentosa (RP) often starts with a normal-appearing retina and only the symptom of nightblindness (or in many cases no symptoms at all when RP is identified early because of family history combined with detection of one of the 300-plus known RP-causing gene mutations). As time passes, retinal arterials in RP show signs of narrowing and being surrounded by melanin-containing cells (the "pigmentosa" part of RP), with underlying areas of retina exhibiting signs of atrophy. Retinal atrophy eventually manifests as reduced visual sensitivity in the corresponding regions of the visual field, which can be measured in the clinic with a test that asks patients to respond when they see a small circle of light of fixed luminosity presented in different parts of the visual field (i.e., kinetic perimetry). As RP continues to progress, the retinal atrophy and other visible pathology spreads and increases in severity. Loss of visual sensitivity in the visual field worsens and increases in area. Eventually central visual acuity is reduced and worsens over time, and in most cases the RP eventually progresses to tunnel vision and in some cases to total blindness. Within the RP population, there is wide variability between individuals in the scaling of various signs and symptoms as a function of patient age. But the similarity between RP patients in the order and dynamics of changes in signs and symptoms is quite robust (Massof & Finkelstein, 1987). The estimation of a latent disease state variable for RP would map the diverse set of observations onto a single continuous measurement scale, thereby extending the range over which the rate of progression of RP can be measured with a single variable.

Chronic open-angle glaucoma (COAG) is another example of an ophthalmic disorder that is defined by a set of signs and symptoms. Anatomically, COAG is characterized by a loss of optic nerve fibers (bundles of axons from retinal ganglion cells) with corresponding pathognomonic changes in the ophthalmoscopic appearance of the optic nerve head. Ganglion cell axons come together to form recognizable patterns of nerve fiber bundles that can be seen on the surface of the retina as they course to the optic nerve head before exiting the eye as the optic nerve. Progressive changes in these anatomical measurements are monitored over repeated visits, both for diagnostic and for treatment decision-making purposes. The nerve fiber damage that occurs in COAG is well studied. It occurs initially at the optic nerve head and then through retrograde degeneration becomes visible in the retinal nerve fiber layer and ultimately progresses to loss of the damaged axons and the retinal ganglion cells from which they originate. The axonal damage at the optic nerve head is presumed to be caused by elevated intraocular pressure (IOP). The only treatment for glaucoma is pharmacologic or surgical intervention that reduces IOP to prevent further damage to nerve fibers. The challenge in managing COAG is that what is defined as abnormally elevated IOP falls within statistical variability of normal IOP, that is, exceeds 2 standard deviations above the normal mean IOP (which necessarily includes 2.5% of the population free from COAG). The ambiguity in the definition of COAG based on IOP is that the effects of IOP on the optic nerve head (the initial location of nerve fiber damage) vary between people. There is a class of glaucoma suspects who have "abnormally" high IOP but no other signs or symptoms of COAG (these patients are identified as having ocular hypertension and are followed as being at risk for nerve fiber damage but are not candidates for treatment).

Low-tension glaucoma is defined as evidence of glaucomatous nerve fiber layer damage despite IOP falling within normal limits.

COAG is often called the "sneak thief of sight." Similar to RP, as glaucoma advances, peripheral vision is lost gradually, starting in the midperipheral visual field and spreading both to the far periphery and toward the center, ultimately resulting in tunnel vision and eventually progressing to total blindness. This patterned vision loss is the only symptom of COAG – there is no pain or discomfort, and visual acuity remains good until later stages of disease progression. Considerable amounts of peripheral vision can be lost before the loss is noticed by the patient. So, by the time the patient reports vision loss, irreversible damage has occurred. To detect early peripheral vision loss, visual sensitivity is measured using standardized test parameters at multiple predefined at-risk visual field locations (i.e., static perimetry) and compared to age-matched norms for the respective location.

Severity of COAG signs and symptoms is typically summarized with ordinal severity ratings of optic nerve damage, average of physical retinal nerve fiber layer thickness measurements at different retinal locations, and the average deviation from age-matched normal means of visual sensitivity measurements at a pre-defined set of visual field locations (called the "mean deviation"), which in turn may be classified as normal, mild, moderate, or severe (Hodapp et al., 1993). The clinical decision to be made from these summary variables, in particular the rate of change in mean deviation or statistically significant changes at individual visual field test locations (e.g., Casas-Llera et al., 2009; Nouri-Mahdavi et al., 2007), is whether or not IOP-lowering treatment should be initiated or modified. Expert consensus is relied upon to guide the interpretation of clinical data and to make COAG management decisions (Weinreb et al., 2011). The estimation of a single latent disease state variable for COAG from the vector of static perimetric visual sensitivity measurements, the vector of nerve fiber layer thickness measurements at different retinal locations, and measurements of anatomical changes in the optic nerve head has the potential of defining a single continuous variable for measuring COAG progression. Such a variable constructed from an array of clinical data can serve to support treatment decision-making and to make measurements of treatment outcomes.

Dry eye is another ophthalmic disorder characterized by a consistent pattern of progression of multiple signs and symptoms, but in this case involving the external eye and adnexa. Dry eye per se is considered a syndrome that can have multiple causes (such as infections, injuries, external or environmental irritants and antigens). Two panels of experts, the National Eye Institute (NEI)/Industry Workshop on Clinical Trials in Dry Eye (Lemp, 1995) and the International Dry Eye Workshops (Lemp et al., 2007), proposed consensus guidelines for differential diagnosis and classifications of etiology: deficient tear production versus abnormal tear evaporation and intrinsic (abnormal anatomy/physiology) versus extrinsic (environmental) causes. Approximately a decade after the NEI/Industry Workshop, a Delphi panel of ocular surface disease experts constructed a set of diagnostic criteria for dry eye, which they renamed "dysfunctional tear syndrome" (DTS). The consensus position that emerged from that study was that DTS should be classified according to the presence or absence of inflammation, with or without signs of eyelid margin disease, and/or abnormal tear distribution and clearance. Although correlations between dry eye signs and symptoms are weak, both the consensus workshops and the Delphi panels agreed that clinicians can estimate the severity of dry eye from the observed signs and symptoms (Behrens et al., 2006). Later studies and theorizing by Baudouin (2007) and Sullivan et al. (2010) concluded that dry eye signs and symptoms are the result of cascading pathophysiological responses to hyperosmolarity of the tear film and that all dry eye cases follow more or less the same course. Sullivan et al. proposed and tested the explicit hypothesis we entertain here that observations and measurements of clinical signs and symptoms can be mapped onto a single dry eye disease state variable (called a "composite variable" by them) that agrees with clinicians' magnitude estimates of dry eye severity. Although their results provided encouraging support of the hypothesis that dry eye signs and symptoms could be mapped onto a single latent dry eye disease state variable, they did not have available to them the analytic metrological tools needed to confirm or refute that specific hypothesis.

Numerous hurdles must be cleared to develop and validate objective measurements of latent disease state variables from vectors of clinically observed manifest variables: the magnitudes of different signs and symptoms are scaled in different units; nonmonotonicity of sign and symptom magnitudes over the course of a disease (i.e., remissions) is plausible in some instances; undoubtedly, there will be large between-person variability in the rate of disease progression; and both within- and between-person variabilities in the dynamics of progression of individual signs and symptoms are likely. However, before addressing these issues we will turn our attention to the required properties, assumptions, and mechanics of latent disease state variables per se, and the magnitudes of which can be estimated from scalable observations of patients' signs and symptoms.

7.3 Theoretical construction of a latent disease state variable

Disease state is a hypothetical scalable trait of the patient. It refers to the stage of progression of the patient's disease, which is defined by the manifestation of a pattern of signs and symptoms. The concept of a latent disease state variable implies that on average the depersonalized defining signs and symptoms of the disease can be placed in a consistent order according to the time at which each is first detected. Each defining sign and symptom may change in magnitude or ordered quality over time after first detection according to its own function, which may vary randomly between patients, expressed in its own units, or at least with ordered categories.

To create a theoretical framework for measurement of disease state (Massof & McDonnell, 2012), we start by defining the continuous latent disease state variable, θ. Theta refers to measurements of the latent variable. When the measurement applies to the latent trait of the *i*th person (P_i), we will denote the measurement as θ_{P_i}. Clinically observed magnitudes or ordered qualities of pathognomonic signs and symptoms are disease state indicators, $I_{i,j}$ for the *j*th sign or symptom of person *i*. The units of $I_{i,j}$ can represent a count, a physical measurement, or an ordinal category (which includes subjective magnitude estimates). Whatever scale the manifest $I_{i,j}$ indicator is on, the quantification of the *j*th observed patient trait can be discretized into $m+1$ ordered categories separated by *m* thresholds. For each indicator variable *j*, the *k*th category is defined as

$$x_{j,k} \le I_{i,j} < x_{j,k+1} \tag{7.1}$$

for which $x_{j,k}$ is in the same units as $I_{i,j}$ and *k* is an integer ranging from 1 to *m*. If $I_{i,j} < x_{j,1}$, then the first category (labeled category 0) is a half-open or open interval from, but not necessarily including, 0 or from, but not including, $-\infty$ at the lower bound up to, but not including, $x_{j,1}$ at the upper bound. Categories 1 to *m* are half-open intervals that include their lower bound, but not their upper bound. Category *m*, for which $I_{i,j} \ge x_{j,m}$, has an upper bound of $+\infty$. The so-assigned ordinal category value of $k_{i,j}$ to the indicator variable value of $I_{i,j}$ must then be converted monotonically to the continuous latent variable units that specify the magnitude of θ_{P_i,I_j,T_k},

$$\theta_{P_i} - \theta_{I_j} - \theta_{T_k} = \theta_{P_i,I_j,T_k} = g_j\left(x_{i,j,k}\right) \tag{7.2}$$

for which $x_{i,j,k}$ specifies the indicator variable threshold for the ordinal category *k* into which $I_{i,j}$ falls; g_j is a monotonic function of the *j*th indicator variable; θ_{I_j} is the value of the *j*th indicator variable in latent variable units; and θ_{T_k} specifies the threshold for ordinal category *k* in latent variable units, irrespective of the indicator variable. Unidimensional latent variable measurement models (Massof, 2011) are used to estimate latent person variable measurements $\left(\theta_{P_i}\right)$ for person *i*, latent indicator variable measurements $\left(\theta_{I_j}\right)$ for indicator *j* (N.B. the latent indicator variable θ_{I_j} is referred to as the “item measure” for “item” *j*), and the estimated measurement of the ranked category threshold *k* on the latent disease state variable scale $\left(\theta_{T_k}\right)$.

Probabilistic models for measurement conjointly estimate values of latent variables from statistical uncertainty in the observed manifest variables. Sources of variability in the estimated variables $\hat{\theta}_{P_i,I_j,T_k}$ include fixed differences between persons, between items, and between thresholds (i.e., fixed differences in the deterministic latent variable θ_{P_i,I_j,T_k}) combined with added stochastic estimation error, ε, i.e.,

$$\hat{\theta}_{P_i,I_j,T_k} = \theta_{P_i,I_j,T_k} + \varepsilon \tag{7.3}$$

for which the expected value of ε is zero and the variance is the expected value of ε^2. The stochastic term, ε, can be assigned to θ_{P_i}, θ_{I_j}, or θ_{T_k}, or apportioned among them.

Models of this kind require that the distribution of ε be the same for all person, indicator variable, and threshold combinations.

Conjoint measurement models take the form of specifying the probability of observing a particular ordinal category, k, given the value of the latent person variable (θ_{P_i}), the magnitude of the indicator variable on the latent variable scale (θ_{I_j}), and the magnitude(s) of the category threshold(s) on the latent variable scale (θ_{T_k}), with k ranging from 1 to m. The measurement model equation for estimating probabilities of observing each ordinal category also requires a specification of the assumed probability density function (PDF) for the random error, ε, in eq. (7.3). Most applications of probabilistic conjoint measurement models assume ε has a logistic distribution with a mean of 0 and a variance of $\pi^2/3$, which closely approximates a standard normal distribution.

Although it is a simple matter to constrain the number of categories, m, to be the same value for all indicator variables, the mapping of I_j to an ordinal category k for each indicator must be more considered. One core assumption here is that statistical uncertainty in the estimated latent variable, $\hat{\theta}_{P_i,I_j,T_k}$, is the same for all observations (i.e., the PDF for ε is identical for all combinations of i, j, and k). Another core assumption is that the latent variable is a monotonic function of the manifest indicator variable assigned to the observation, $g_j(x)$.

A priori, we do not know what the empirical distributions of observed $I_{i,j}$ values will be for a sample of patients. However, the expectation of probabilistic conjoint measurement models is that the distribution of θ_{P_i} values in a patient sample is independent of the indicator variables and their category thresholds, and the distribution of the deviate ε is the same for all values of θ_{P_i}, θ_{I_j}, and θ_{T_k}.

A deterministic system has no randomness – changes of state are perfectly predictable. A stochastic system has varying degrees of randomness manifesting as unpredictable deviations from an expected deterministic value or other fixed reference. Entropy refers to the degree of randomness in a system or set of observations. The more uniform the distribution of deviations of observed values from expected values, the greater is the entropy.

Entropy and information are two sides of the same coin, but colloquially they are often described as opposites (e.g., information is desirable, entropy is undesirable; information is true, entropy is false; information is message, entropy is corruption; information is signal, entropy is noise). From this qualitative viewpoint, one might consider a discretization strategy that minimizes entropy to be optimal. The problem with the colloquial characterization of entropy and information as opposite is the underlying assumption that information is equivalent to knowledge, whereas within the context of measurement, information is just encoded data that are distributed in some fashion. At the level of individual observations, the state of ignorance is high – samples that would be categorized by an omniscient observer as information is confusable with samples that would be categorized by the omniscient observer as entropy. Both encoding of information and encoding of entropy depend on resolution. Observations that can resolve entropy also can resolve information. Therefore, an encoding scheme that maximizes observed entropy is also expected to maximize observed information (i.e., Jaynes's principle of maximum entropy) (Jaynes, 1957).

The observed indicator variable distribution for a sample of subjects has two sources of entropy: within-subject randomness (i.e., test-retest variability) and between-subject randomness (stochastically distributed but fixed differences between subjects). Encoding of observations is accomplished by creating ordinal categories or equal interval bins for each indicator variable (resolution is defined by the number of bins used to encode the range of observations). The indicator variable is observed for each subject, recorded, and added to the count in the corresponding ordered category or equal interval bin. The total count in each bin is normalized to the number of subjects to estimate the relative frequency of each indicator value in the form of a histogram, that is, a probability mass function (PMF). If all the observations were to fall in the same category or bin, entropy (and information) would be zero. At the other extreme, if the frequency of observations was the same for all categories or bins, then entropy (and information) would be maximized. If bins were created with quantiles instead of equal intervals, then the frequency distribution would be uniform (rectangular) on an ordinal equal frequency (quantile) scale. Therefore, the optimal number of quantiles for binning data to maximize the information encoded corresponds to the number that maximizes entropy.

A priori, each of N samples of an indicator variable has the same $1/K$ prior probability of falling into any one of the K equal interval size bins. Considering all possible outcomes ranging from all samples falling in the same bin to all bins having the same number of samples, there are

$$W = \frac{N!}{\prod_{k=1}^{K} N_k!} \tag{7.4}$$

possible permutations of N indicator variable values sorted into K bins with N_k samples in each bin, that is,

$$\sum_{k=1}^{K} N_k = N \tag{7.5}$$

The natural log of a factorial is

$$\ln(N!) = \sum_{n=1}^{N} \ln(n) \tag{7.6}$$

which when substituted for log factorials in the log of eq. (7.4), the log of the number of permutations is

$$\ln(W) = \sum_{n=1}^{N} \ln(n) - \sum_{k=1}^{K} \sum_{j=1}^{k} \ln(j) \tag{7.7}$$

For large N (>25 for <5% error), ln(N!) can be approximated as

$$\ln(N!) \approx N \ln N - N \tag{7.8}$$

which is Stirling's approximation. Using Stirling's approximation, eq. (7.7) is reexpressed as

$$\ln(W) \approx N \ln N - N - \sum_{k=1}^{K} (N_k \ln N_k - N_k) \tag{7.9}$$

for which N_k is the number of samples in bin k. Defining the probability of a sample falling in bin k to be $p_k = N_k/N$, and working through the algebra, we conclude that

$$\ln(W) \approx -N \sum_{k=1}^{K} p_k \ln(p_k) \tag{7.10}$$

which is the formula for Gibbs entropy (without the Boltzmann constant). Normalizing ln(W) to N, we obtain an expression of Shannon's entropy,

$$H = \frac{\ln(W)}{N} = -\sum_{k=1}^{K} p_k \ln(p_k) \tag{7.11}$$

The units of Shannon entropy (uppercase eta) are determined by the base of the logarithm ("nats" for the natural log, "bits" for $\log_2$ – the unit used most in information theory). Shannon entropy, H, is proportional to the expected value of information in the K bins. The maximum possible Shannon entropy that could be observed for a uniform distribution of the indicator variable (K equal interval bins with $p_k = 1/K$ for all intervals) is

$$H_{\max} = -\ln\left(\frac{1}{K}\right) \tag{7.12}$$

which is the same as Shannon's information.

Entropy in a single bin is $H_k = -p_k \ln(p_k)$. Entropy adds, so total entropy for K bins is simply the sum of H_k for $k = 1$ to K, as defined in eq. (7.11). Two factors determine entropy: (1) the number of bins that contain observations (K) and (2) the relative distribution of observations across the available bins (i.e., distribution of p_k). From eq. (7.11), we can see that empty bins ($p_k = 0$) do not contribute to entropy (or to information). (N.B. If $p_k = 0$, then $H = -0 \times \ln(0) \equiv 0$)

To maximize entropy (H in eq. (7.11)) when our knowledge of the distribution is limited to an estimate of the mean, we add the constraint that the expected value of indicator variable $I_{j,k}$ is

$$\mathbb{E}\{I_{j,k}\} = \sum_{k=1}^{K_j} p_{j,k} I_{j,k} \tag{7.13}$$

for K_j discrete $I_{j,k}$ intervals. We also add the constraint for all indicators that the probability of falling in one of the mutually exclusive and exhaustive K_j discrete intervals is 1:

$$\sum_{k=1}^{K_j} p_{j,k} = 1 \tag{7.14}$$

Employing the method of Lagrange multipliers to maximize entropy, we define

$$\mathcal{L}_j = -c\sum_{k=1}^{K_j} p_{j,k}\ln\left(p_{j,k}\right) - \beta\sum_{k=1}^{K_j} I_{j,k}p_{j,k} + \beta\mathbb{E}\left\{I_{j,k}\right\} - \alpha\sum_{k=1}^{K_j} p_{j,k} + \alpha \tag{7.15}$$

At maximum entropy,

$$\frac{\partial\mathcal{L}_j}{\partial p_{j,k}} = -c\ln\left(p_{j,k}\right) - c - \beta I_{j,k} - \alpha = 0 \tag{7.16}$$

which results in the expression

$$p_{j,k} = e^{-\left(1+\frac{\alpha}{c}\right)}e^{-\left(\frac{\beta I_{j,k}}{c}\right)} \tag{7.17}$$

an exponential PDF. Exponential PDFs, $\lambda e^{-\lambda x}$, have one parameter, λ. The mean of an exponential PDF is $1/\lambda$ and the variance is $1/\lambda^2$. In the first exponential on the right-hand side of eq. (7.17), α/c is an empirical constant, as is β/c, which weights the indicator variable $I_{j,k}$ in the second exponential. With this understanding, we conclude

$$\lambda = e^{-\left(1+\frac{\alpha}{c}\right)} = \frac{\beta}{c} \tag{7.18}$$

in eq. (7.17), which means that at maximum entropy $c = \beta/\lambda$. Maximum entropy for an exponential distribution is

$$H_{\max} = 1 - \ln(\lambda) \tag{7.19a}$$

which, after substitution of terms for an exponential PDF with mean $1/\lambda$, is expected to be

$$H_{\max} = 2 + \frac{\alpha}{\beta}\lambda \tag{7.19b}$$

If the variance of the PDF is independent of the mean, then another constraint is added to the Lagrange multiplier in eq. (7.15):

$$\gamma\mathbb{E}\left\{I_{j,k}^2\right\} - \gamma\sum_{k=1}^{K_j} p_{j,k}I_{j,k}^2 = 0 \tag{7.20}$$

which results in

$$p_{j,k} = e^{-\left(1+\frac{\alpha}{c}\right)} e^{-\left(\frac{\beta I_{j,k} + \gamma I_{j,k}^2}{c}\right)} \tag{7.21}$$

a class of distributions that includes the normal distribution, for which the variance equals σ^2 and

$$H_{\max} = 0.5\left(1 + \ln\left(2\pi\sigma^2\right)\right) = 0.5 + \ln\left(\sigma\sqrt{2\pi}\right) \tag{7.22}$$

For a normal distribution with a mean (μ) equal to 0, $c = 2\sigma^2$, $\alpha = c\ln(\sigma\sqrt{2\pi}) - c$, $\beta = -2\mu = 0$, and $\gamma = 1$ in eq. (7.21); therefore,

$$H_{\max} = 1.5 + \frac{\alpha}{c} \tag{7.23}$$

Most applications of probabilistic conjoint measurement models employ a version of the model that assumes the uncertainty PDF, in units of the latent variable θ, is a logistic distribution for which the variance is $\sigma^2\pi^2/3$ and maximum entropy is

$$H_{\max} = 2 + \ln(\sigma) \tag{7.24}$$

If the PDFs of the manifest indicator variables are unknown, but empirically the number of trials is the same in every bin k (equal frequency intervals of the indicator variable for the same distribution of the latent disease state variable θ_{P_i}, i.e., distributions of all indicator variables are constructed from observations on the same group of subjects), entropy (and information) for K_j equal frequency bins is maximized for each observed distribution (re. eq. (7.12)). Maximum entropy for a fixed number of bins does not require the bins to have the same dimensions on an equal interval scale, they only must be equal frequency (i.e., $p_{j,k}$ defines a constant quantile of the indicator variable distribution – it has the same value for every bin of every indicator variable).

To divide each manifest indicator variable into m equal uncertainty intervals for a sample intended to represent the targeted patient population, each interval for each indicator is expected to have the same entropy (i.e., the same degree of randomness). The left panel of Figure 7.1 shows the nonmonotonic mathematical relationship between probability and Shannon's entropy in nats (i.e., between $p_{j,k}$ and H in eq. (7.11)). The consequence of equating entropy is that a unique ordinal scale is defined for each indicator variable in I_j units that satisfies the requirement of converting $I_{j,k}$ to an equal interval scale in entropy units. Because entropy adds (see eq. (7.11)), entropy increases as the number of bins increases. The right panel of Figure 7.1 shows the relationship between maximum entropy and the number of equal frequency bins (quantiles) ranging from dichotomization of the distribution of observations (2 bins, each with a probability of 0.5 and an entropy of 0.347 nats) to conversion of the distribution of observations to percentiles (100 bins, each with a probability of 0.01 and an entropy of 0.046 nats).

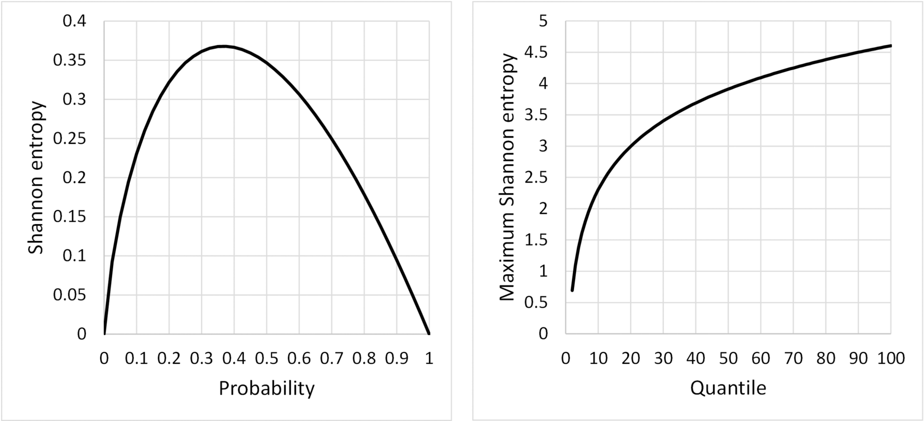


Figure 7.1: (Left panel) Shannon's entropy in "nats" (natural log units) as a function of probability. The probability at which maximum entropy occurs depends on the base of the logarithm (0.5 for $\log_2$ and 0.368 for the natural log). (Right panel) Maximum Shannon entropy in nats as a function of the number of equal frequency bins (quantiles) ranging from 2 (dichotomized at the median) to 100 (percentiles).

This is not the first time this class of measurement models has been linked to entropy (Pendrill, 2019; Pendrill et al., 2019; Melin et al., 2021; Melin & Pendrill, 2023). Indeed, Melin et al. (2021) concluded that the dichotomous conjoint scaling model "can be viewed as an entropy-based measurement model." Their use of entropy focused mainly on information carried by probabilistic conjoint measurement models expressed in the form of a reduction of entropy owing to conditional probabilities for $p_{j,k}$ (expressed as the change in entropy when conditioned on the observations). Our approach is based on the assumption that measurement information is maximized when observed entropy is maximized, which occurs when the observed indicator variable PMFs on an equal entropy scale are approximately rectangular.

7.4 Estimation of dry eye disease state

To demonstrate the application of these theoretical concepts to the estimation of a latent disease state variable, we analyzed dry eye data generously shared with us by Benjamin Sullivan and his co-workers. Sullivan et al. (2010) measured dry eye signs and symptoms in each eye of 299 subjects, approximately half of whom were dry eye patients and the other half normal volunteers. Measurements of dry eye indicator variables for each subject eye included tear osmolarity (in mOsm/L), tear volume (Schirmer's test expressed as mm of wetting of a filter paper wick making contact with the eye under the lower lid), tear film breakup time (TBUT – in seconds post blink), meibomian gland dropout score (Foulkes & Bron, 2003), corneal staining-type score (NEI/Industry

Workshop fluorescein staining-type scoring system), corneal staining area score, conjunctival staining-type score (NEI/Industry Workshop scoring system), conjunctival staining area score, and Ocular Surface Disease Index (OSDI) rating scale questionnaire raw score (patient self-report of symptom severity) (Schiffman et al., 2000). Indicator variable values have either a positive monotonic relationship with dry eye severity (e.g., tear osmolarity increases as dry eye worsens) or a negative monotonic relationship with dry eye severity (e.g., TBUT decreases as dry eye worsens). Except for the OSDI, measurements of the dry eye indicator variables were made on the right and left eyes of each subject separately. These raw dry eye data were collected, analyzed, and reported earlier by Sullivan et al. (2010). That study was repeated by our group several years later on 203 dry eye patients and 51 normal controls (Karakus et al., 2018), but the results were analyzed using the multiplicative polytomous model described by Masters (1982). To illustrate the application of these models to the estimation of latent disease state variables, we report here the results of applying a polytomous logistic difference model (Bradley & Massof, 2018) to the raw data of Sullivan et al. (2010) after rebinning their data by converting them to an equal entropy scale.

For each of the nine dry eye indicator variables reported in Sullivan et al.'s (2010) study, $I_{i,j}$ for $j = 1$ to 9, we computed the empirical relative frequency of each observed indicator variable value, $p(I_{i,j})$, from the recorded measurements for each eye of each of the 299 subjects. As demonstrated in Figure 7.2 by the cumulative distributions of indicator variable measurements for the left eye (OS – solid curves) and for the right eye (OD – dashed curves) of dry eye patient subjects (blue curves) and of normal subjects (red curves), average indicator variable values of normal subjects are less than the average value of dry eye patient subjects for seven of nine indicator variables, the reverse is true for tear breakup time and for Schirmer's test. There is good agreement between right and left eyes (re: superposition of solid and dashed curves with Pearson correlations ranging from 0.19 for tear osmolarity in normal subjects and 0.42 in dry eye subjects to 0.97 for meibomian gland dropout in normal subjects and 0.93 in dry eye subjects). The two eyes cannot be compared for the OSDI raw scores because subjects were not instructed to apply their ratings of items to each eye separately.

Because the main source of information for constructing measurements in this context is test-retest variability, we treated each observation of each of the indicator variables for each eye of each subject as an independent observation (i.e., latent variable models of this type assume local independence). Figure 7.3 illustrates histograms of the relative frequency distributions of the indicator variables as a function of each variable's magnitude for each eye tested, expressed in its own measurement units on either an equal interval physical measurement scale (tear osmolarity through Schirmer's test in Figure 7.3) or ordinal rating scale (corneal staining type through OSDI in Figure 7.3). To assign values on a common scale for the diverse array of observed indicator variables for each subject eye, we employed Shannon's entropy (Shannon, 1998) as the common unit of measurement. Shannon's entropy, eq. (7.11), is a measurement of the amount of uncertainty in the random indicator variable $I_{j,k}$, irrespective of

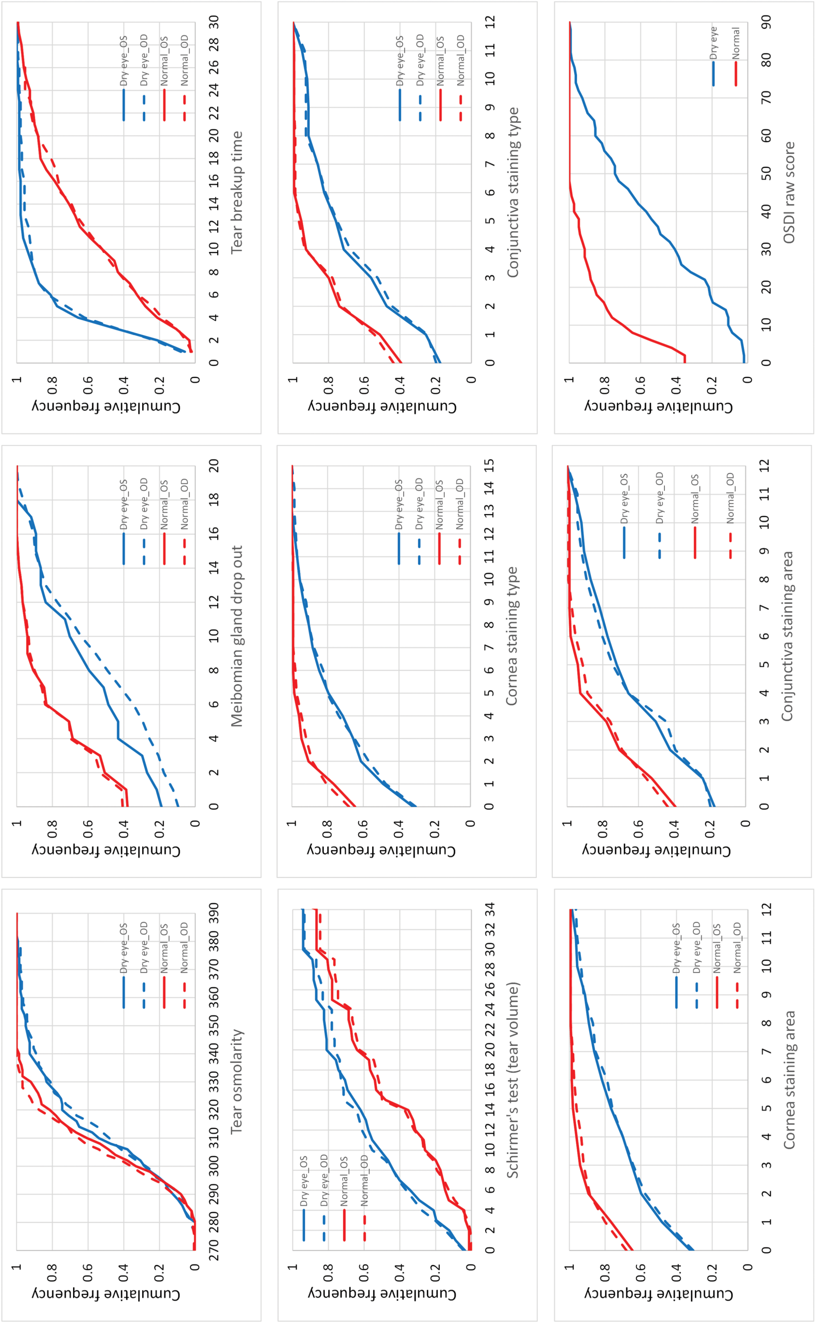
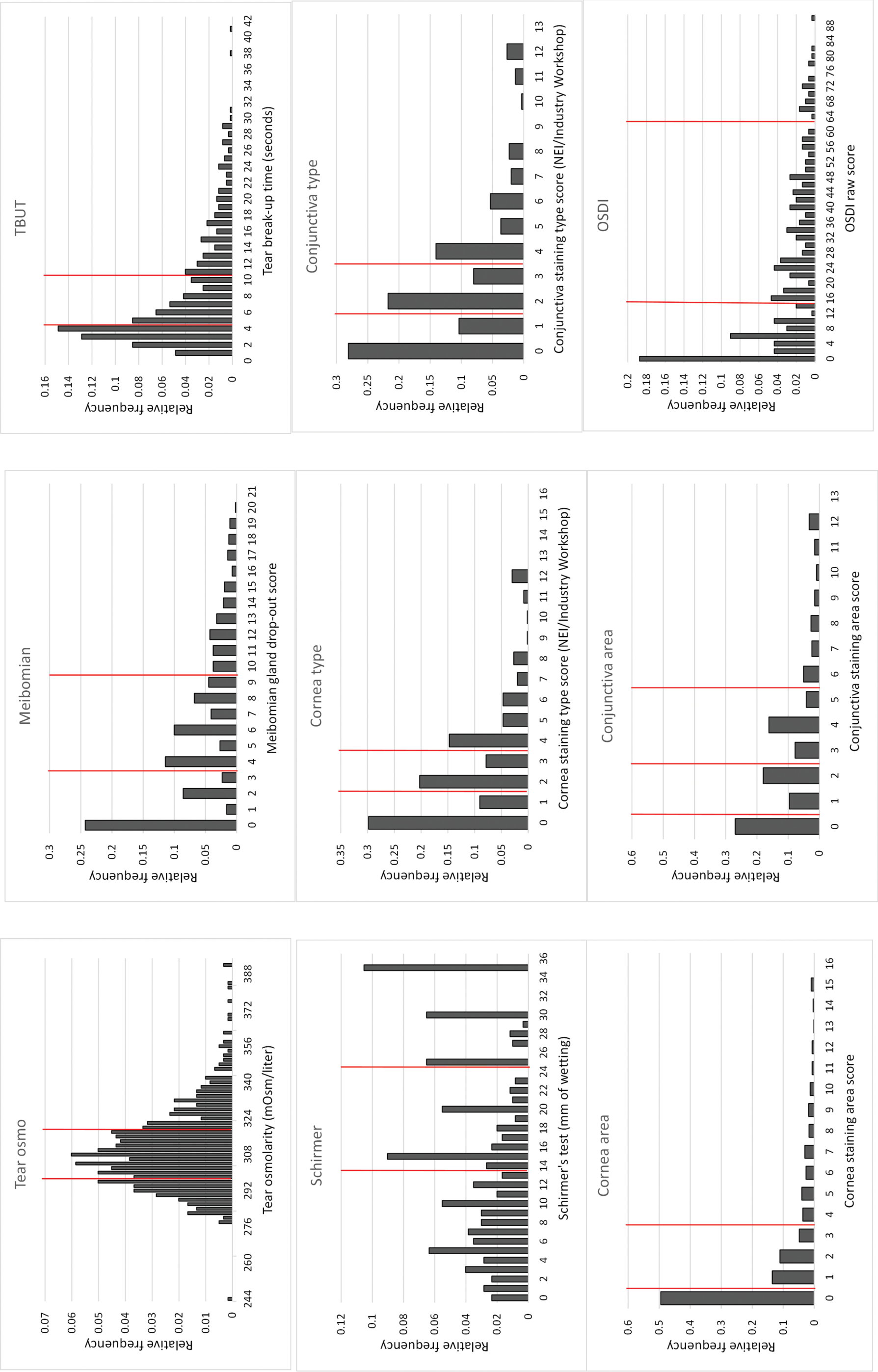


Figure 7.2: Cumulative frequency distributions of nine different dry eye indicator variables (panels) in left eyes (solid curves) and right eyes (dashed curves) of dry eye patients (blue functions) and normal control subjects (red functions).


TBUT
Relative frequency
Tear break-up time (seconds)
Conjunctiva type
Conjunctiva staining type score (NEI/Industry Workshop)
OSDI
OSDI raw score
Meibomian
Meibomian gland drop-out score
Cornea type
Cornea staining type score (NEI/Industry Workshop)
Conjunctiva area
Conjunctiva staining area score
Tear osmo
Tear osmolarity (mOsm/liter)
Schirmer
Schirmer's test (mm of wetting)
Cornea area
Cornea staining area score

units in which the variable is expressed, over the range of a specified interval *k* for the *j*th indicator (here expressed in “nats” because log base *e* is employed):

$$H_{j,k} = -p\left(I_{j,k}\right) \times \ln\left[p\left(I_{j,k}\right)\right] \tag{7.25}$$

Each of eight of the nine indicator variable distributions (Figure 7.3) is divided into three intervals (defined by red vertical lines) with approximately the same Shannon entropy in each interval ($H_{j,k} \cong 0.36$ nats). The size of the intervals was chosen to match the maximum observed entropy of all the bins used to record observations of all the indicator variables (which in Sullivan et al.’s data is 0.361 nats for the *cornea-type* indicator variable). The distribution of *conjunctival staining areas* was divided into four intervals, the first three with approximately the same entropy (0.36 nats) as each of the three intervals for the other eight indicator variables, the fourth interval with the leftover entropy ($H_{9,4} \cong 0.30$ nats). Entropy adds (see eq. (7.11)), so total entropy for each indicator variable ranges from 1.04 to 1.09 for the first eight indicators (interval 4 had zero entropy for those eight indicators) and is 1.37 for the ninth indicator (see Figure 7.4). So even though each indicator variable is recorded in different units (*x*-axis for each graph in Figure 7.3), some on an ordinal scale and others on an equal *interval* scale, the discretization of observations into approximately equal intervals on an entropy scale maximizes the likelihood that uncertainty will be the same for each scored observation irrespective of the differences in the original indicator variable units. By scoring observations on an approximately equal entropy scale, we satisfy the fundamental homogeneity of variance assumption of this class of measurement models.

Each histogram in Figure 7.3 illustrates the relative frequency distribution in the sample of 598 eyes of each of the nine indicator variables (designated on the abscissa). The red vertical lines in each graph denote the threshold values that divide the indicator variable into three (first eight distributions for which the three equal entropy intervals spanned all available data) or four approximately equal entropy intervals (conjunctival stain area distribution, which had leftover entropy after the entropy in the first three intervals was equated with entropy in the three intervals for the other indicators). The first interval is always from 0 up to, but not including, the left-most

Figure 7.3: (continued)
Combined relative frequency distributions (probability mass functions) for right and left eyes of each dry eye and normal subject. Each subject eye is treated as an independent observation for each of the nine indicator variables (displayed in separate panels). In each panel, the red vertical lines are category thresholds for each of the indicator variables. Category 0 ranges from 0 to the first red line, category 1 ranges from the first to the second red line, and category 2 ranges from the second red line to the end of the scale. The ordinal category labels are reversed for the Schirmer test and tear breakup time distributions. Each category defines a range of the indicator variable over which the entropy is approximately 0.36 nats. The exception is for conjunctival stain area, for which the first three categories have an entropy of about 0.36 and a third threshold defines a fourth category with leftover entropy of about 0.3 nats.

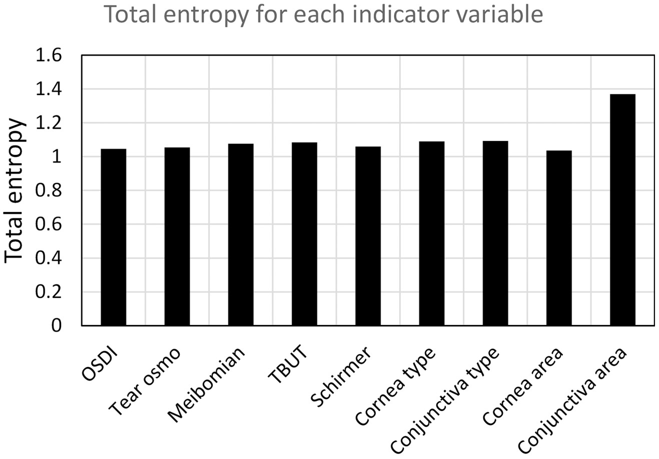


Figure 7.4: Comparisons between dry eye indicator variables of total entropy, that is, sum of entropy in each of three equal entropy intervals (for the first eight indicator variables) or four entropy intervals (for the ninth indicator variable for which the first three intervals have the same entropy as the other indicator variable intervals and the fourth interval has the leftover entropy).

red line; the second interval is from the left-most line up to, but not including, the second line; and the last interval is from the right-most line to the end of the scale.

An equal entropy interval value (0, 1, 2, or 3) for each of the 9 indicator variables is recorded for each of the 598 subject eyes studied (for a possible total of 5,382 unique combinations of subjects' eyes and dry eye indicators). The scale was reversed for the two indicator variables that monotonically decrease with dry eye severity (Schirmer's test and TBUT). There were missing values for 10% of all possible subject eye/indicator pairings in the Sullivan et al. database we employed (more than half of the missing values can be attributed to the OSDI questionnaire being administered once to each subject and the results assigned to only one of the subject's eyes, the other missing values are conjunctival staining area scores for the right eye).

Using the method of successive dichotomizations (Bradley, 2021), we employed a polytomous logistic difference model to estimate θ_{P_i} for each person, θ_{I_j} for each indicator variable, and θ_{T_k} for each equal entropy interval threshold ($k = 1$ separates intervals 0 and 1, $k = 2$ separates intervals 1 and 2, and for conjunctival stain area, $k = 3$ separates intervals 2 and 3). Person measures could not be estimated for 11 subject eyes, because scores were at the minimum for all indicators, and for 9 subject eyes because scores were at the maximum for all indicators. The top panel in Figure 7.5 is a histogram of the estimated person measures (i.e., distribution of θ_{P_i} – estimated latent dry eye disease state for each of the 577 subject eyes that were included in the analysis). The bottom panel in Figure 7.5 is a histogram of the estimated item (dry eye indicator) plus threshold measures (i.e., distribution of $(\theta_{I_j} + \theta_{T_k})$ – estimated for each sum of item measure and $k = 1$ and $k = 2$ equal entropy threshold measures for all items plus $k = 3$ for only the conjunctival stain area item, yielding a total of 19 $(\theta_{I_j} + \theta_{T_k})$ values displayed in Figure 7.5. It can be appreciated that the distribution of $(\theta_{I_j} + \theta_{T_k})$ values in the lower panel is well-centered on the distribution of θ_{P_i} values in the upper panel.

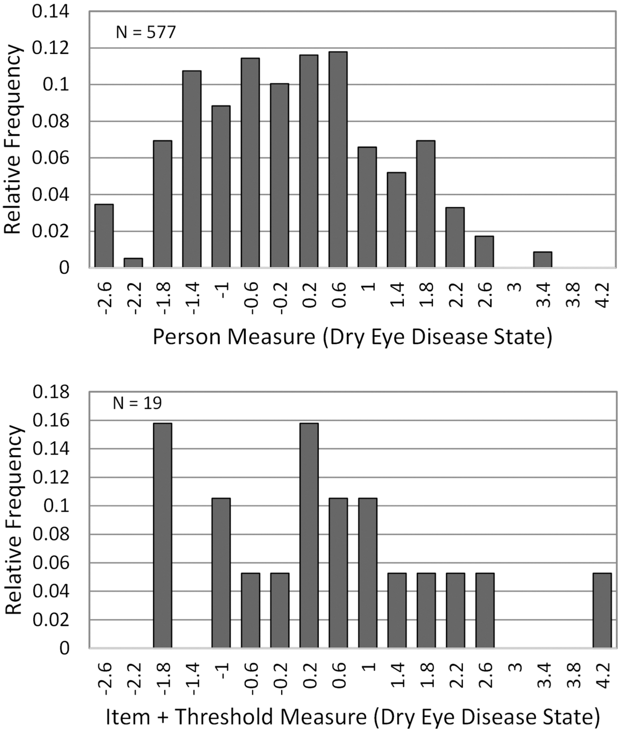


Figure 7.5: (Top panel) Histogram of person measures (θ_{P_i}) estimated for 577 of 598 subject eyes. Estimates could not be made for 21 eyes that had extreme ordinal entropy scores (all 0 or all 3 or 4) for all indicator variables for which there were data for the eye. (Bottom panel) Histogram of the sums of the 9 estimated item measures and 2 (or 3 for the conjunctival staining area item) category thresholds $(\theta_{I_j} + \theta_{T_k})$.

Because entropy in the raw observed indicator values, $I_{i,j}$, was constrained to be approximately the same for all intervals of all indicator variables, irrespective of the units or distribution of the observed manifest variables, each estimated latent dry eye disease state item measure and person measure is expected to have approximately the same statistical resolution. This expectation is the consequence of a core inferential requirement of this class of measurement models that uncertainty is uniform on the continuous latent variable scale. That is, for a completely deterministic model, measurements would have no random errors (no uncertainty) and the PDF for repeated measurements would be a Dirac delta function (i.e., a single probability value of 1 at that point on the latent variable axis).

However, if repeated measurements resulted in a distribution of errors around a mean point on the latent variable axis that could be characterized by a PDF, for example, a Gaussian or logistic distribution with a mean equal to $\theta_{P_i} - \theta_{I_j} - \theta_{T_k}$ and a standard deviation equal to $\sigma_{i,j,k}$, then the uniform distribution of probabilities on the latent variable axis would be the consequence of a convolution of the uncertainty PDF with each point on the latent variable axis (obviously, even under ideal conditions, there would be departures from a uniform distribution of integrated probabilities when approaching the endpoints of the range of measurements). Uncertainty of each estimated measure can be summarized with the standard error of the estimate. The left panel of Figure 7.6 displays the standard error of each person measure estimate as a function of the estimated person measure and the right panel displays the standard error of each item measure estimate as a function of the estimated item measure.

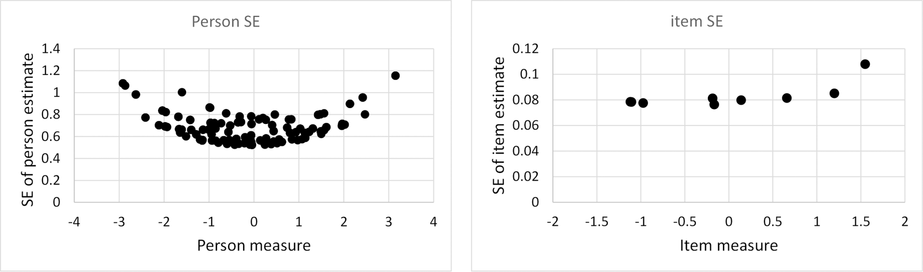


Figure 7.6: (Left panel) Scatter plot of the standard error of the person measure estimate versus the estimated person measure for 577 subject eyes (note there are large numbers of overlapping points). (Right panel) Scatter plot of the standard error of the item measure estimate versus the estimated item measure for the nine indicator variables.

Note that the standard error of the item measure is about an order of magnitude less than the standard error of the person measure. That difference can be attributed to the larger number of persons contributing to each item measure estimate ($N = 577$) than the number of items contributing to the person measure estimate ($N = 7–9$). Since the standard error of the estimate can be regarded as the standard deviation of the estimation error (SD$_e$) normalized to the square root of the degrees of freedom (1 less than the number of observations that contributed to the estimate), the SE for person measures should be approximately 8 times larger than the SE for the item measures if the SD$_e$s are the same for both item and person measure estimates. We computed the SD$_e$ for each item by multiplying that item measure's SE by the square root of its degrees of freedom (number of persons contributing observations to the item measure estimate less than 1). Similarly, we computed the SD$_e$ for each person by multiplying that subject eye's person measure SE by the square root of its degrees of freedom (number of items that contributed observations to that subject eye's person measure less than 1). Figure 7.7 shows the resulting scatterplot of person measure SD_e values for each subject eye (red points – note that there is heavy overlap of the 577 points) and of item measure SD_e values for each of the 9 indicator variables (blue points – 2 of the points overlap) along with the horizontal line at $SD_e = 1$, the expected value of σ used in probabilistic conjoint measurement models (the consequence of the specific constant chosen for the SD_e is a compensatory change in the size of the dimensionless logit unit employed for the estimated measurements). The systematic increase in the SD_e from 1 at the extremes is the consequence of increasing contributions to the estimated measurements of unbounded extreme item measure intervals (i.e., the measurement scale extends to negative infinity on the left and to positive infinity on the right).

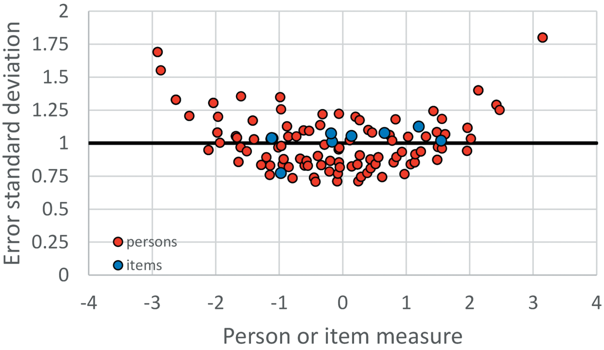


Figure 7.7: Scatter plots of the estimated standard deviations of the person measure uncertainty distribution for each person versus the estimated person measure (red points) and the estimated standard deviations of the item measure uncertainty distribution for each of the nine items (blue points). The standard deviations were calculated by multiplying the standard error of the estimate for each person and item in Figure 7.6 by the square root of its degrees of freedom. The solid line is at the expected value of the standard deviation ($\sigma = 1$) of ε in eq. (7.3).

Measurements are conjoint and noninteractive. The conjoint part of measurement comes from the comparison of a trait of the entity to be measured (*measurand*) to a measurement scale (*measurement instrument*). The measurement instrument not only defines the scale but also specifies the measurement operations. For measurements to be noninteractive, the measurement instrument cannot alter the measurand and the measurand cannot alter the measurement instrument. Literally, volumes have been written on axiomatic measurement theory (Krantz et al., 1971; Suppes et al., 1989; Luce et al., 1990), but most formal measurement theories ignore measurement error and effectively assume that measurement systems are deterministic (Michell, 1999). The genius of probabilistic conjoint measurement models is that they enable equal interval scales constructed from uniform stochastic measurement error (Perline et al., 1979). However, estimated measurements from these models are expected values that must conform to a deterministic conjoint structure that is known as a Guttman scale (Massof, 2004).

The ordinal scores assigned to observations for each pairing of an item and person are entered into the corresponding cell of N persons × J items matrix (left panel of Figure 7.8). If the rows are ordered by estimated person measures and the columns are ordered by estimated item measures, then if the ordinal scores are monotonic with the estimated measures, $\hat{\theta}_{P_i,I_j,T_k}$, they should have an ordered progression for each row and for each column, thereby defining a Guttman scale. Ordinal scores representing equal entropy intervals (see Figure 7.3) for each combination of subject eye (row) and dry eye indicator variable (column) are entered into the appropriate cell of the matrix in the left panel of Figure 7.8 and color-coded (0, blue; 1, green; 2, yellow; 3, red; and missing data, white). The 9 columns in the matrix have been ordered by estimated item measures and the 577 rows have been ordered by estimated person measures.

The matrix in the right panel of Figure 7.8 illustrates the color-coded ordinal scores expected by probabilistic conjoint measurement models with the estimated latent disease state variables θ_{P_i} and $\theta_{I_j} + \theta_{T_k}$. To test the goodness of fit of observed results from equal entropy interval transformations of dry eye indicator variable values in the left panel to model expectations of dry eye indicator values from the estimated person and item measures in the right panel, we can employ a normalized chi-square statistic called the information-weighted mean square fit statistic – a.k.a. “infit” (N.B. “information-weighted” refers to Fisher information, not to Shannon’s information). The person infit for each subject eye is simply the sum across columns of squared differences be-

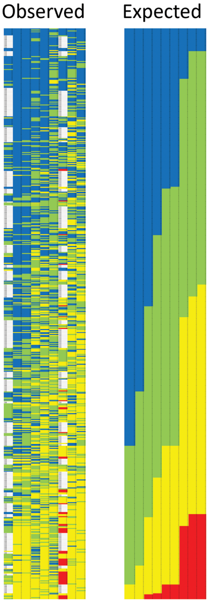


Figure 7.8: (Left panel) Guttman scalogram of scores assigned to equal entropy intervals that contained the observed indicator variable value for each combination of person (rows) and indicator variable (columns). (Right panel). Expected Guttman scalogram for the measurement model with the estimated person measure for each subject eye, the estimated item measure for each indicator variable, and estimated threshold measure for each category threshold. For both scalograms, the scores for each person/item combination are color coded (blue, 0; green, 1; yellow, 2; red, 3).

tween the observed value (left matrix) and corresponding expected value in that subject eye's row (right matrix) divided by the degrees of freedom for the subject eye (i.e., number of cells in the subject eye's row, less than 1, that have observed values). As evidenced by the white cells in the left panel, approximately 10% of subject eyes have fewer than 8 degrees of freedom. Therefore, the observed infit distribution across subject eyes is expected to be a weighted sum of χ^2/df distributions.

Since probabilistic conjoint measurement models are axiomatic and have no free parameters, rather than testing the fit of model expectations (right panel of Figure 7.8) to the observations (left panel of Figure 7.8), we test the fit of the observations to the expectations of the axiomatic measurement model. If the fit falls within statistical resolution limits, we conclude that the model assumptions are satisfied by the observations and the estimated person and item measures are valid measurements within error tolerances.

Figure 7.9 displays the histogram of person measure infits for the 577 subject eyes. The red curve plotted along with the data is the expected weighted sum of χ^2/df PMFs corresponding to the distribution of the number of indicator variables with data for

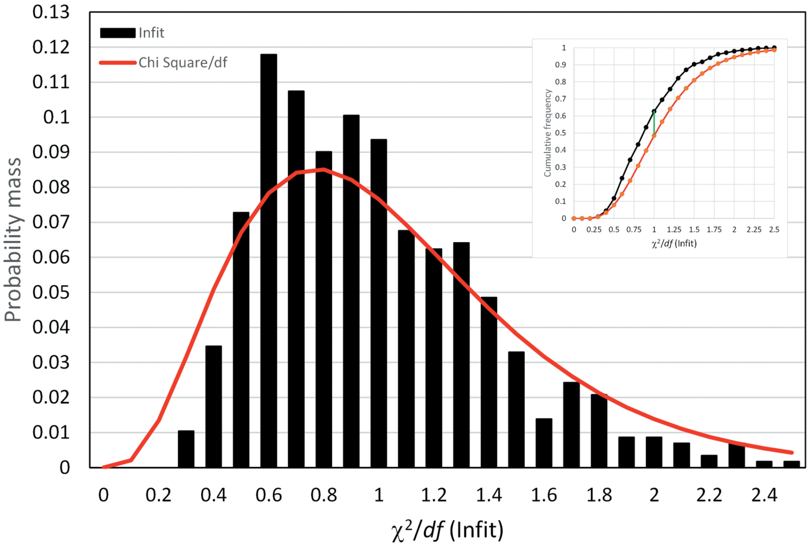


Figure 7.9: Histogram of the information-weighted mean square fit statistic (infit) for estimated person measures (black bars). The person measure infit is the sum across indicator variables of the squared difference between the observed indicator variable entropy bin score and the score expected by the measurement model, given the estimated person, item, and threshold measures, divided by the degrees of freedom (number of indicator variable observations for the person minus 1). The red curve is the χ^2/df PMF predicted by a weighted mixture of χ^2/df PMFs, with weights corresponding to the fraction of persons having each of the observed degrees of freedom. The inset shows the cumulative distribution functions for the observed infit (black) and expected (red) PMF. The Kolmogorov-Smirnov test (not significant) was applied to the cumulative functions.

each subject. (N.B. The expected value of χ^2 is its degrees of freedom, so the expected value of χ^2/df is 1. Degrees of freedom is the only parameter defining a χ^2 distribution. The χ^2 distribution is skewed for small df: the mean = df, the variance = 2df, the median = df $(1–2/9\text{df})^3$, and the mode is the maximum of df – 2 or 0, so the median and the mode of χ^2/df → 1 as df → ∞. Wilson and Hilferty (1931) showed that the distribution of $(\chi^2/\text{df})^{1/3}$ is well approximated by a normal distribution when df > 25.) From the Kolmogorov-Smirnov test we conclude that there is no significant difference between the cumulative distributions (inset in Figure 7.9) for the observed infits (black histogram bars) and for the weighted sum of χ^2/df PMFs (red curve). To increase statistical resolution of the fit, it would be necessary to increase the number of items (indicator variables) and/or increase the number of thresholds for each indicator variable, thereby increasing the degrees of freedom in estimating the person measure and reducing the standard error of person measure estimates.

Although we do not necessarily expect bilateral symmetry in dry eye disease severity, Figure 7.7 shows strong agreement between eyes (solid vs dashed lines) in the indicator variable cumulative distribution functions for both dry eye and normal subjects. Consistent with those results, the scatter plot in the left panel of Figure 7.10 shows strong agreement between estimated dry eye severity person measures for the right and left eyes relative to the red identity line (Pearson's correlation = 0.83). Figure 7.7 also shows that all indicator variable cumulative frequency distributions for dry eye subjects are shifted relative to distributions for normal subjects (blue vs red lines). Consistent with those results, the cumulative distribution functions in the right panel of Figure 7.10 show that the latent dry eye disease severity distribution is shifted to greater values for dry eye subjects (blue curve) relative to the distribution for normal subjects (red curve). These results are consistent with the findings of Karakus et al.'s (2018) replication of the Sullivan et al. (2010) study (see figure 7A in Karakus et al. (2018)).

7.5 Interpretation of analysis results

Patient-centered outcomes can be classified according to the methods employed: (1) patient-reported outcomes (PROs); (2) clinician-reported outcomes (ClinRO); (3) observer (or proxy)-reported outcomes (ObsRO); and (4) performance outcomes (PerfO) (Cano et al., 2019). The difference between ClinRO and ObsRO is the type of knowledge required of the observer to judge the outcome being rated (specialized clinical knowledge of signs and symptoms vs personal knowledge of the patient's preferences, behavior, and history). Both clinicians (ClinRO) and proxies (ObsRO) report ranks assigned to their judgments about their personal observations of the patient, whereas patients (PRO) report their personally ranked responses to questions or statements (items) that they apply to themselves. PerfO consist of measurements of physical clinical variables. In Sullivan et al.'s (2010) dry eye study, the OSDI results are considered PROs, the cornea

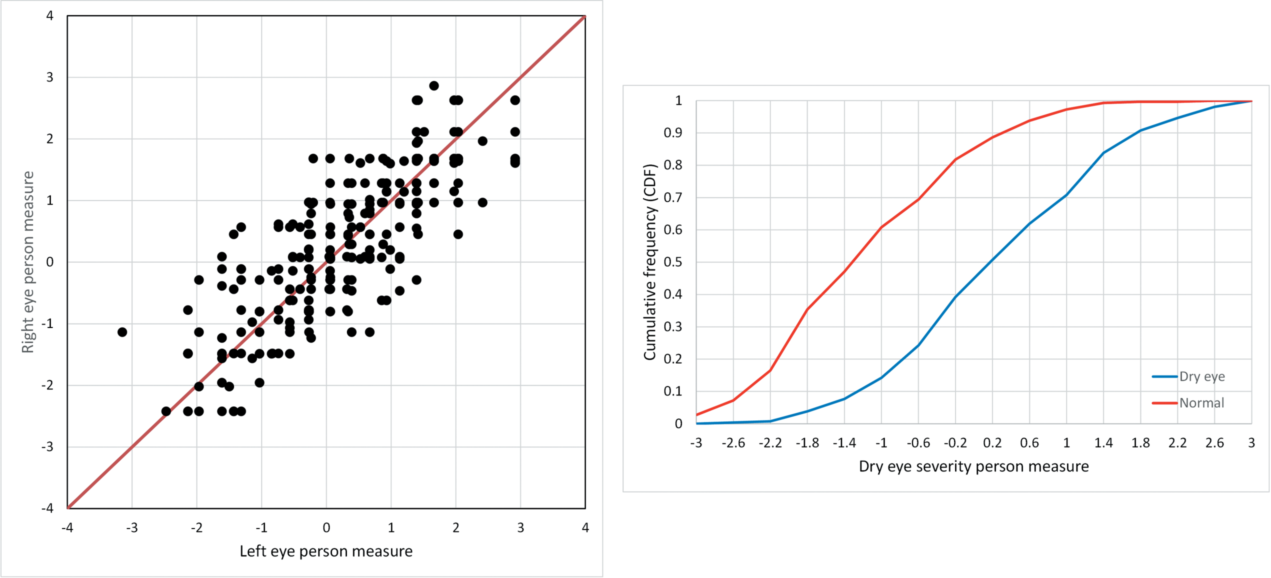


Figure 7.10: (Left panel) Scatterplot of estimated person measures for the right eye versus person measures for the left eye of all normal and dry eye subjects combined (circles). The solid red line is the identity line. The Pearson correlation is 0.83. (Right panel) Cumulative person measure distribution functions for normal subject eyes (red curve) and for dry eye subject eyes (blue curve). The two curves are separated by 1.6 logits ($p < 0.001$; two-tailed t-test).

and conjunctival type of staining and area of staining are considered ClinROs, and tear osmolarity, meibomian gland dropout count, tear breakup time, and Schirmer's test are considered PerfOs. To estimate a single dry eye disease state variable for each patient from this diverse array of observations requires that the magnitudes of the different observed variables be expressed in the same units. We have employed a strategy of defining the common unit for all observations to be entropy (here expressed as "nats"). (N.B. By definition, quantiles are intervals that have the same probability and therefore have the same entropy. The reverse is not true because of the nonmonotonic relationship between entropy and probability. As shown in the left panel of Figure 7.1, each entropy value below the maximum corresponds to two different probability values.)

A "unit" of measurement has a fixed size even though it may be infinitely divisible. Equal size units are added (or concatenated) until a match with the measurand is achieved. The resulting measurement corresponds to a count of the number of units required for a match. But measurements are only as accurate and repeatable as the accuracy and repeatability of the observations. Repeated measurements produce a distribution of observations with respect to an expected value (usually the mean). One specification of measurement accuracy is "bias," which refers to the difference between the mean of repeated measurements and the "true" or "correct" value. The evaluation of bias requires a reference standard against which the measurement instrument is calibrated, and the true value of the measurement is defined (Pendrill, 2019). Another form of measurement accuracy is the validity of the assumptions implicit in the structure of the measurement instrument (e.g., measuring distance on a line vs measuring distance on a curve) (Massof, 2010). The precision of a measurement refers to the unit size (e.g., the number of significant figures in the recording of the measurement) (Pendrill, 2019). Precision can also refer to the repeatability of the measurements per se (e.g., standard deviation of the repeated measurements distribution, as discussed relative to Figure 7.7) or to the repeatability of, or confidence in, the mean measurement (i.e., standard error of the mean, as discussed relative to Figure 7.6) (Pendrill, 2019).

By converting all raw observations, both ordinal ClinROs and PROs and interval-scaled PerfOs, to the same equal entropy units for all indicator variables, we effectively have made the differences between frequency distributions for the different indicator variables (re: Figure 7.3) irrelevant. Those frequency distributions are sample-dependent (i.e., dependent on the selection of dry eye patients and normal subjects), whereas the functions that convert each of the indicator variables to a common latent dry eye disease state variable (eq. (7.2)) must be monotonic and sample-independent (i.e., the person trait or property is the measurand and the items, thresholds, and PDF of the random deviate ε constitute the noninteracting measurement instrument).

Probabilistic conjoint measurement models are axiomatic, their construction defines the measurement rules – they have no free parameters. The simplest such model is dichotomous – it has a single threshold added to the item measure that divides the θ_{P_i, I_j, T_k}scale into two intervals, scored $k = 1$ for observations above the item measure plus threshold and $k = 0$ for observations below the item measure plus threshold. If we were to dichotomize the observations in Sullivan et al.'s dry eye study, we would place the sum of the single threshold and item measure for each indicator variable at the median of its PMF in Figure 7.3 (creating two equal entropy intervals of 0.347 nats for each indicator variable). It is important to note that the threshold for each indicator variable is estimated a posteriori. Thus, there is uncertainty in the repeatability of indicator variable PMFs for the same sample of persons, let alone uncertainty in PMF repeatability between samples of persons. Uncertainty in the conversion of equal entropy intervals of observed indicator variable values to θ_{P_i, I_j, T_k} is modeled by the PDF chosen for ε. The most commonly employed dichotomous model assumes the PDF for ε is a logistic distribution with a mean of 0 and a standard deviation of 1.

Birnbaum's dichotomous Item Response Theory (IRT) model also uses a logistic PDF for ε_j, but allows the standard deviation of the PDF, σ_j (the reciprocal of which is called the item discrimination parameter), to be estimated separately for each indicator variable (Birnbaum, 1968). Allowing the standard deviation of ε_j to vary across items is equivalent to creating local distortions (stretching and compression) of the latent variable measurement scale around different items (i.e., changing the size of the measurement unit), which would require a far more complex ad hoc theory of the latent variable to justify and explain the distortions (analogous to the metric tensor in general relativity theory used to describe the curvature of space-time distortions from gravity or analogous to adding local changes in elevation to measurements of distance on a contour map).

We employed a logistic difference version of the dichotomous model (Bradley & Massof, 2018) for our analysis of the Sullivan et al. dry eye data. The logistic difference model is identical to Samejima's (1969) graded response IRT model except that a logistic PDF is used for ε instead of a Gaussian PDF, a benign difference, and the item discrimination parameter is fixed to a constant value of 1 instead of being a free parameter that is estimated for each indicator variable. In earlier attempts to construct a dry eye disease state metric (Karakus et al., 2018), the partial credit model (Masters, 1982), a polytomous measurement model first described in a general form by Rasch (1961) and later developed by Andersen (1973, 1977), was used to estimate θ_{P_i, I_j, T_k}. The partial credit model, which is built on the assumption that the thresholds are statistically independent, is a multiplicative model – the product of dichotomous models for each threshold (Massof, 2011). The partial credit model also employs a logistic PDF and fixes σ in the ε distribution to the same constant for all indicator variables.

A problem raised with applications of the partial credit model is that estimates of category thresholds, θ_{T_k}, can be disordered (i.e., $\theta_{T_k} < \theta_{T_{k-1}}$) (Bradley and Massof, 2018; Jansen & Roskam, 1986; Roskam & Jansen, 1989; Luo, 2005; Massof, 2012). Tutorials suggest that estimates of disordered thresholds can be attributed to a problem with the data (Bond & Fox, 2015). However, mathematically the partial credit model imposes no constraints on threshold ordering – in effect thresholds are dissociated from the concatenated intervals they are expected to define. For the model to describe a measurement scale, one must assume that the *m* thresholds are always ordered, they are not statistically independent random variables. That assumption is equivalent to assuming that the thresholds separating ordered categories are identified by their order after the fact (viz., by definition, threshold *k* separates category *k*-1 from category *k*), not that the thresholds are separate entities that are identified and permanently labeled a priori and then tracked as they randomly and independently wander trial-to-trial in their positions on the latent variable scale.

The polytomous measurement model, as conceptually formulated by Rasch (1961) and further developed by Andersen (1973, 1977), takes the form of a conditional probability driven by the requirement of *statistical sufficiency of raw scores* (Andersen, 1977). The person raw score is the sum of ordinal values for person *i* ($k_{i,j}$) assigned to each indicator variable, $S_i = \sum_{j=1}^{J} k_{i,j}$, and the item raw score is the sum of ordinal values for item *j* ($k_{i,j}$) assigned to each person, $S_j = \sum_{i=1}^{N} k_{i,j}$. Statistical sufficiency of the raw score means that S_i is sufficient to estimate θ_{P_i} and S_j is sufficient to estimate θ_{I_j}. In the case of dichotomous scoring ($m = 1$), there is only one threshold θ_{T_k} and statistical sufficiency results in the production of a Guttman scale. In the case of polytomous scoring ($m \geq 2$), adherence to a Guttman scale of the pattern of scores in a N person × J item matrix requires statistical sufficiency of raw scores, but statistical sufficiency of raw scores does not require adherence to a Guttman scale. To satisfy the axioms of measurement theory, all measurements must conform to a Guttman scale (noninteractive conjoint structure). Therefore, the constraint must be added to the model that thresholds are ordered and define ordered intervals (Bradley & Massof, 2018).

The assumptions of the partial credit model can be satisfied if category thresholds are regarded as “steps” that must be taken in a specified order, not as boundaries defining ordered concatenated intervals (Masters, 1982; Luo, 2005; Massof, 2012; Wright & Masters, 1982). Like steps taken in solving a math problem, the position of a step on the latent variable scale refers to its difficulty, whereas its identifying ordinal value refers to the order in which it must be completed. With this interpretation, thresholds are not boundaries between concatenated intervals, they are hurdles of varying height that must be cleared in a specified order.

7.6 Conclusions

These results, as well as those of similar past studies of dry eye (Massof and McDonnell, 2012; Karakus et al., 2018), reinforce the application of probabilistic conjoint measurement models to the estimation of latent disease state variables from scalable patient signs and symptoms obtained on their own scales (whether ordinal values or a variety of equal interval physical measurement units). Our two previous dry eye studies employed the partial credit version of a multiplicative polytomous model (Masters, 1982), the analysis described here employed the polytomous logistic difference version of the dichotomous model (Bradley & Massof, 2018). When data conform to the modeled expectations with no disordered thresholds in the partial credit context, the two models estimate person and item measures that are linearly related (they differ in scale) (Bradley & Massof, 2018). The largest difference between these two types of measurement model is in how thresholds are defined and estimated (Massof & Bradley, 2023) and whether conditional probabilities (multiplicative models) or unconditional probabilities (logistic difference models) are employed. Both the present analysis of Sullivan et al.'s (2010) data, and the earlier Karakus et al.'s (2018) study converted the raw indicator variable values recorded in different units to an ordinal common scale before performing the measurement scaling analysis. The present analysis discretized the diverse indicator variable values by defining a common scale as equal entropy intervals; Karakus et al.'s study defined the common scale as quintiles of indicator variable values, which is equivalent to an equal entropy scale because each quintile has an entropy of 0.322.

To the extent that the presence of clinical disorders is inferred from an array of signs and symptoms, each scaled in its own units, but are first detected and progress in magnitude according to their own time-dependent functions, we can use probabilistic conjoint measurement models to employ all observed clinical information to estimate a single latent disease state variable. Such a variable provides a quantitative outcome measurement – one suitable for inclusion in an extended scheme of SI units – that uses all relevant clinical observations to study the natural history of the disease and measure the efficacy of interventions.

References

Andersen, E. B. (1973). Conditional inference for multiple choice questionnaires. British. *Journal of Mathematical and Statistical Psychology*, *26*, 31–44.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*, 69–81.

Baudouin, C. (2007). Un nouveau schema pour mieux comprendre les maladies de la surface oculaire. *Journal Francais d'Ophtalmologie*, *30*, 239–246.

Behrens, A., Doyle, J. J., Stern, L., et al. (2006). Dysfunctional tear syndrome: A Delphi approach to treatment recommendations. *Cornea*, *25*, 900–907.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.). *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. 3rd ed., New York, NY USA: Routledge.

Bradley, C. (2021). msd: Method of successive dichotomizations, https://CRAN.R-project.org/package=msd.

Bradley, C., & Massof, R. W. (2018). Method of successive dichotomizations: An improved method for estimating measures of latent variables from rating scale data. *PLoS ONE*, *13*(10), e0206106.

Cano, S. J., Pendrill, L. R., Melin, J., & Fisher, W. P., Jr. (2019). Towards consensus measurement standards for patient-centered outcomes measurement. *Measurement*, *141*, 62–69.

Casas-Llera P, Rebolleda G, Muñoz-Negrete FJ, Arnalich-Montiel F, Pérez-López M, Fernández-Buenaga R. (2009). Visual field index rate and event-based glaucoma progression analysis: comparison in a glaucoma population. *British Journal of Ophthalmology, 93*. (12):1576–1579. doi: 10.1136/bjo.2009.158097.

Foucault, M. (1973). *The birth of the clinic: An archaeology of medical perception*. [Translation of *Naissance de la Clinique* (1963) by A.M. Sheridan Smith], New York, NY USA: Pantheon Books.

Foulkes, G. N., & Bron, A. J. (2003). Meibomian gland dysfunction: A clinical scheme for description, diagnosis, classification, and grading. *The Ocular Surface*, *1*, 107–126.

Hodapp, E., Parrish, R. K., II, & Anderson, D. R. (1993). *Clinical Decisions in Glaucoma. St* (pp. 52–61). Louis, MO USA: CV Mosby Company.

Jansen, P. G. W., & Roskam, E. E. (1986). Latent trait models and dichotomization of graded responses. *Psychometrika*, *51*, 69–91.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, *106*(4), 620–630.

Karakus, S., Akpek, E. K., Agrawal, D., & Massof, R. W. (2018). Validation of an objective measure of dry eye severity. *Translational Vision Science and Technology*, *7*, 26. https://doi.org/10.1167/tvst.7.5.26

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement. I. Additive and polynomial representations*. San Diego, CA and London, UK: Academic Press.

Lemp, M. A. (1995). Report of the National Eye Institute/Industry workshop on clinical trials in dry eye. *Contact Lens Association of Ophthalmologists Journal*, *21*, 221–232.

Lemp, M. A., Baudouin, C., Baum, J., et al. (2007). The definition and classification of dry eye disease: Report of the definition and classification subcommittee of the international dry eye workshop (2007). *The Ocular Surface*, *5*, 75–92.

Luce, R. D., Krantz, D. H., Suppes, P., & Tversky, A. (1990). *Foundations of measurement. III. Representation, axiomatization, and invariance*. San Diego, CA and London, UK: Academic Press.

Luo, G. (2005). The relationship between the rating scale and partial credit models and the implication of disordered thresholds of the Rasch model for polytomous responses. *Journal of Applied Measurement*, *6*, 443–455.

Massof, R. W. (2002). The measurement of vision disability. *Optometry and Vision Science*, *79*, 516–552.

Massof, R. W. (2004). Likert and Guttman scaling of visual function rating scale questionnaires. *Ophthalmic Epidemiology*, *11*, 381–399.

Massof, R. W. (2010). A clinically meaningful theory of outcome measures in rehabilitation medicine. *Journal of Applied Measurement*, *11*, 253–270.

Massof, R. W. (2011). Understanding Rasch and item response theory models: Applications to the estimation and validation of interval latent trait measures from responses to rating scale questionnaires. *Ophthalmic Epidemiology*, *18*, 1–19.

Massof, R. W. (2012). Is the partial credit model a Rasch model? *Journal of Applied Measurement*, *13*, 114–131.

Massof, R. W., & Bradley, C. (2023). An adaptive strategy for measuring patient-reported outcomes: Incorporating patient preferences relevant to cost-benefit assessments of vision rehabilitation. In W. P. Fisher Jr. & S. J. Cano (Eds.). *Person-centered outcome metrology.* Springer Series in Measurement Science and Technology. Cham: Springer, https://doi.org/10.1007/978-3-031-07465-3_5

Massof, R. W., & Finkelstein, D. (1987). A two-stage hypothesis for the natural course of retinitis pigmentosa. *Advances in Biosciences*, *62*, 29–58.

Massof, R. W., & McDonnell, P. J. (2012). Latent dry eye disease state variable. *Investigative Ophthalmology and Visual Science*, *53*, 1905–1916.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Melin, J., Cano, S., & Pendrill, L. (2021). The role of entropy in construct specification equations (CSE) to improve the validity of memory tests. *Entropy*, *23*, 212–227.

Melin, J., & Pendrill, L. R. (2023). The role of construct specification equations and entropy in the measurement of memory. In W. P. Fisher Jr. & S. J. Cano (Eds.). *Person-centered outcome metrology:*

Principles and applications for high stakes decision making. Cham: Springer Series in Measurement Science and Technology. Springer, https://doi.org/10.1007/978-3-031-07465-3_102023

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ USA: Lawrence Erlbaum Associates.

Michell, J. (1999). *Measurement in psychology. A critical history of a methodological concept*. Cambridge, UK: Cambridge University Press.

Nouri-Mahdavi, K., Hoffman, D., Ralli, M., Caprioli, J. (2007). Comparison of methods to predict visual field progression in glaucoma. *Archives of Ophthalmology, 125*. (9):1176–1181. doi: 10.1001/ archopht.125.9.1176.

Pendrill, L. (2019). *Quality assured measurement: Unification across social and physical sciences*. Springer Nature Switzerland AG.

Pendrill, L. R., Melin, J., Cano, S. J., et al. (2019). Metrological references for healthcare based on entropy. 19th International Congress of Metrology, 07001, https://doi.org/10.1051/metrology/201907001

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*, 237–255.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In: J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 321–334). Berkeley CA USA: University of California.

Roskam, E. E., & Jansen, P. G. W. (1989). Conditions for Rasch-dichotomizability of the unidimensional polytomous Rasch model. *Psychometrika*, *54*, 317–333.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, *34*(2), 17.

Schiffman, R. M., Christianson, M. D., Jacobsen, G., Hirsch, J. D., & Reis, B. L. (2000). Reliability and validity of the ocular surface disease index. *Archives of Ophthalmology*, *118*, 615–621.

Shannon, C. E. (1998). The mathematical theory of communication (1948). In C. E. In: Shannon & W. Weaver (Eds.). *The mathematical theory of communication* (pp. 29–116). Urbana and Chicago, IL USA: University of Illinois Press.

Stegenga, J. (2018). *Care and cure: introduction to philosophy of medicine*. Chicago, IL USA: The University of Chicago Press.

Sullivan, B. D., Whitmer, D., Nichols, K. K., Tomlinson, A., Foulks, G. N., Geerling, G., Pepose, J. S., Kosheleff, V., Porreco, A., & Lemp, M. A. (2010). An objective approach to dry eye disease severity. *Investigative Ophthalmology and Visual Science*, *51*, 6125–6130.

Suppes, P., Krantz, D. H., Luce, R. D., & Tversky, A. (1989). *Foundations of measurement. II. Geometrical, threshold, and probabilistic representations*. San Diego, CA and London, UK: Academic Press.

Weinreb, R. N., Garway-Heath, D. F., Leung, C., Crowston, J. G., & Medeiros, F. A. (2011). *Progression of Glaucoma: 8th Consensus report of the World Glaucoma Association*. Amsterdam, The Netherlands: Kugler Publications.

Wilson, E., & Hilferty, M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences U.S.A*, *17*, 684–688.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL USA: MESA Press.

W. Steve Lang and Judy R. Wilkerson

8 Measuring teacher dispositions: steps in an innovative journey in affective assessment

Abstract: The affective domain (dispositions), as opposed to the cognitive domain (knowledge and skills or performance), in teacher education is not well assessed in most cases despite the requirement to do so in the national standards for educator preparation programs. Attention in teacher education is devoted almost exclusively to being "competent" over "caring." In this chapter, an innovative approach composed of nine steps in assessing dispositions is presented. Examples from a 15-year research agenda testing a battery of assessments entitled "Dispositions Assessments Aligned with Teacher Standards" are provided. The accreditation requirement is detailed, along with each of the nine steps and five instruments used. Measurement properties are discussed from various previous studies. The approach is recommended for other professions in which human interaction between alumni and client is important. The repeated demonstration of persistent structural invariances points toward potentials for productively extending the SI.

Keywords: affective measurements, teacher dispositions, accreditation, high stakes assessment

8.1 Introduction

Researchers have found that there is a strong correlation between teacher dispositions and students' learning (Bradley et al., 2020), and measurement of those dispositions has been a long and continuing important part of the preparation program (Phelps, 2006; Choi et al., 2016). All teacher preparation programs in the USA that seek national accreditation are required to demonstrate that their graduates have acquired the Interstate Teacher Assessment and Support Consortium (InTASC) identified knowledge, skills, and dispositions promulgated by the Council of Chief State School Officers (2013). However, while teacher educators are adept at instructing and assessing teacher knowledge and performance, and while certification tests on knowledge and skills are subject to oversight by measurement professionals, the affective domain typically falls by the wayside,

W. Steve Lang, Department of Educational Measurement, College of Education, University of South Florida, St. Petersburg, Florida, USA

Judy R. Wilkerson, Department of Leadership, Research, and Technology, College of Education, Florida Gulf Coast University, Florida, USA

https://doi.org/10.1515/9783111036496-008

treated informally as an afterthought or without systematic attention to standards-based decision-making, taxonomic interpretation of results, or well-conceived measurement techniques. The typical mainstream approach of using observations, Likert items, and confirmatory factor analysis has proven highly unsatisfactory (Niu et al., 2017).

In this chapter, we share our experiences in measuring teacher dispositions (the affective domain) in the accreditation context. We present the nine-step process model we used to develop, test, and use an innovative battery of assessments called *Dispositions Assessment Aligned with Teacher Standards* (DAATS). We note that most affective assessment procedures in the teacher education literature are based on a single instrument that includes a relatively random selection of items loosely tied to professional standards.

The four-pronged idea undergirding all of our work on dispositions assessment – that has held constant over time – is that (1) items should be written to show consistency with professional standards; (2) multiple instruments of different item types are necessary; (3) that design and scoring must be tied to a meaningful taxonomy that defines levels in the scale; and (4) advanced measurement practice provides a solid foundation for use of the results in scoring candidates while seeking to improve both their individual performance as well as the performance of the program. Building on results obtained in research on socio-emotional learning (Pancorbo et al., 2021; Lei et al., 2023), by mapping variation in the measurements in terms of the item content, the process we describe helps us in diagnosing opportunities for individual and program improvements and to avoid construct under-representation and construct irrelevant variance (Baghaei, 2008; Stenner & Horabin, 1992; Wilson, 2023).

As our first innovation/prong, we begin with a standards-based design applied to a battery of five assessments (second innovation). There is attention in the literature to adapting various forms of dispositions, mostly linked to professional behaviors or personality characteristics (Bradley et al., 2020) and to use a research base and a solid definition of the term “disposition” for that development (Rike & Sharp, 2008). What is lacking in the literature, however, are instruments that are tightly aligned with the InTASC standards of teaching (Lang et al., 2020; Wilkerson & Lang, 2007; Wilkerson, 2012). In our work, every item in every assessment is aligned with the national teaching standards (InTASC Standards), developed under the auspices of the Council of Chief State School Officers (CCSSO, 2013), and the battery uses multiple measures of different formats.

As our third innovative approach, we applied the Krathwohl Affective Taxonomy (Krathwohl et al., 1964), the second phase of the Bloom *Taxonomy of Educational Objectives* (Bloom, 1956). Bloom’s taxonomy (1956) has been applied for decades in educational assessment, and in our work, we apply an affective taxonomy to all the items for design, scoring, and interpretation. Items in each assessment were developed to reflect taxonomic levels as well as the InTASC Standards; constructed response or observational items were scored based on the Krathwohl taxonomy.

Fourth, we applied probabilistic models of measurement science (Andrich, 1988; Bond & Fox, 2007; Bond et al., 2021; Rasch, 1980) to analyze and validate the items and measurements. Commonly referred to as "the Rasch model," this family of models has well-known identities with the Bradley-Terry-Luce and Zermelo models for paired comparisons (Andrich, 1988, p. 43; Linacre, 1995, 2000) and conforms to the principles of additive conjoint formulations of fundamental measurement (Newby et al., 2009; Smith et al., 1994; Wright, 1997). Finally, metrologists, the physicists, and engineers who manage the SI units, commonly known as the metric system of measurement, are expressing increasing interest in probabilistic measurement models of this kind, recognizing that they "belong to the same class that metrologists consider paradigmatic of measurement" (Mari & Wilson, 2014, p. 326; also see Mari et al., 2023; Pendrill, 2019; Pendrill & Fisher, 2015). There are, then, strong reasons for both considering the possibility of extending the SI and for broadening the conceptual scope of probabilistic conjoint measurement beyond the narrow confines of a singular association with Rasch. Situating his remarkable contributions in the long history of equivalent contributions deserving of recognition and use has the added benefit of making those contributions more accessible to those who may balk at attributing an entire sphere of modes and methods developed by many contributors over several generations to only one person.

There are excellent examples of the utilization of these models in taxonomic scoring in cognitive assessment (Mohamed et al., 2008) but not for the affective domain. In the pages that follow, we review the evolution of the standards we have applied and then go back to the beginning of our work in 2007, so we can present our journey. We summarize our experiences and results, as we continue to develop a meaningful approach to the heightened focus in teacher education accreditation on teacher dispositions in addition to teacher competence. As such, the innovative steps we have taken, along with our sharing of sample applications and results, have potential for application in any discipline which needs metrics for both competence and caring, especially if such metrics are required for accreditation. We begin with our accreditation context as teacher educators.

8.2 The accreditation context in teacher education

8.2.1 Recognizing teacher dispositions as an important construct: the beginning

Decades ago, Katz and Raths (1986) suggested that the goals of a teacher education program should include not only the acquisition of knowledge and skills but also a class of outcomes they proposed to call "dispositions." They suggested that the construct be defined in terms of description and classification – "the trend of a teacher's actions across similar contexts" (p. 2). They were the first to suggest that dispositions become a criterion for competence and that undesirable dispositions be used as a criterion for disqualifying a candidate. Over the decades, we have learned that the measurement of teacher dispositions can be useful for program improvement in colleges of education (LaPaglia, 2020; LaPaglia & Wilkerson, 2023), which is a fundamental requirement in accreditation (Wilkerson, 2019).

In this early thinking, Katz and Raths (1986) proposed several procedures including both intuitive and measurement approaches. For the latter, they suggested the selection of a panel of experts to nominate acts and categories, for example, "to be enthusiastic," with specific examples to be observed. Enthusiasm could be measured by the teacher's use of appropriate non-verbal signals, use of various tones, or showing delight. A second panel would rate the nominations, with the highest ranked nominations entered on a scale to be judged by observers and then summed to arrive at an "enthusiasm" score. They proposed that the results could be used both to evaluate the candidates as well as the program's efficacy; however, the linkage to accepted measurement techniques (both classical and modern test theory) was not made.

Frustration set in with the identification of specific measurable dispositions, which they feared would become as troublesome as it had been with the competency-based education efforts of the past. Teaching and personality were difficult to separate. Compounding that difficulty were attributions in which one observed or classified a behavior positively while another might classify the same behavior as "insensitive." Other approaches that were tried to define the construct remain a challenge.

Katz and Raths (1986) are largely responsible for subsequent work over the decades that recognized the importance of teacher dispositions and classifying dispositions as a construct to be measured – one that was soon to become required of teacher candidates graduating from accredited American teacher preparation programs. However, the road to construct definition was not easy, and for some years, debate raged over the underlying basis for that definition. Were the dispositions of teachers to be tied to morality or to something else? We briefly cover that debate next.

8.2.2 The morality versus standards debate

In earlier years, the focus in teacher dispositions was focused more on morality-based assessment (Wilkerson, 2006), but the more detailed the InTASC focus became, the more institutions shifted toward explicit professional standards. While we advocated for using standards as the basis for assessment, Dottin (2009) placed the definition of the dispositions construct in morality as exhibited through "habits of mind" based on "moral knowledge." He wrote:

> Identifying, nurturing and assessing habits of mind in professional education programs presuppose that candidates will transfer their learning to the world of practice. Accordingly, working

> toward getting candidates, and faculty members in professional education programs to demonstrate habits of mind through wisdom in practice is a moral endeavor in terms of doing the right thing at the right time for the right reason with the right people (Phelan, 2001, 2005), for "what is learned and employed in an occupation having an aim and involving cooperation with others is moral knowledge, whether consciously so regarded or not." (Dewey, 1944, p. 356)

Many ascribed to the moral basis for assessing teacher dispositions, and this philosophy is well articulated by Diez (2007), with her roots in a faith-based private college. Borko et al. (2007) continued the debate over the ethical and moral basis for dispositions, concluding that NCATE (1987) had unleashed something out of control:

> Given all this attention and controversy, it appears that NCATE's goal to initiate conversations about the moral and ethical development of teachers was achieved. But what, exactly, has been unleashed? Was this move on the part of NCATE brilliance or folly? And perhaps most centrally, do dispositions have a place in the professional standards for teachers or programs to prepare teacher candidates? (p. 359)

For us, the answer is now, and always has been, in the standards themselves. A search on "moral" in the InTASC Standards yields the word "moral" only once in all of the critical dispositions. The critical disposition that "The teacher practices the profession in an ethical manner," has one statement at the highest level of performance that reads: "The teacher collaborates with colleagues to deepen the learning community's awareness of the moral and ethical demands of professional practice." The word "ethical," however, appears 31 times including in the standard (#9) of *Professional Learning and Ethical Practice*.

Britannica defines the terms, noting that although they are often used interchangeably, generally the distinction is that morality tends to have a Christian connotation while ethics is the term used more in conjunction with the professions in reference to a personal code of conduct for workers in those fields. The debate can be endless. In our work, however, in the year 2007 when the debate was raging in the literature, we published our two books on *Measuring Teacher Competency* and *Measuring Teacher Dispositions* – both bearing the second component of the title: "Five Standards-Based Steps Using the CAATS (or DAATS) Model." Our philosophy has been consistent over the years – follow the standards. It is how we avoid what Borko et al. (2007) questioned as brilliance or folly.

8.2.3 Accreditation of teacher preparation programs: the evolution of the national requirement to assess teacher dispositions

Concurrent with the release of the Katz and Raths (1986) call for measuring teacher dispositions, the National Council for Accreditation of Teacher Education (NCATE) redesigned its standards to provide for a more meaningful review process, which they termed the “NCATE Redesign” (Wilkerson, 1987). The 1987 Standards began with *Design of the Curriculum* which included professional practice, ethics, and culturally diverse and exceptional populations. There were no specific standards to support these aspects of the curriculum – just “honorable mention.”

As the accreditation standards and processes evolved, dispositions took on a greater role first in the NCATE standards, as they were revised, and subsequently in the work of the merger of NCATE with its competitor, the Teacher Education Accreditation Council (TEAC) in 2013 into a new agency, the Council for Accreditation of Educator Programs (CAEP), which exists today. In its current glossary, CAEP defines dispositions as “the habits of professional action and moral commitments that underlie an educator’s performance,” citing the InTASC Model Core Teaching Standards (p. 6), as their own source.

In 1992, shortly after Katz and Raths (1986) had called for the assessment of teacher dispositions as a construct and after the accreditation standards and process were redesigned, a definition of dispositional traits emerged from the Council of Chief State School Officers (CCSSO) in 1992. That year CCSSO’s InTASC released its teacher standards, which incorporated and defined the knowledge, skills (performances), and dispositions required of beginning teachers. The InTASC standards included separate expectations for knowledge, performance, and critical dispositions for each of the 10 Standards. At long last, the dispositions for teachers were identified with a high degree of specificity. They generally focus on how and to what extent the teacher values the knowledge and skill relevant to the standard. We often quip that if we spend many hours teaching a candidate to plan, and the candidate does show the skill well but hates it, the candidate is likely not to plan well in the classroom. If they don’t like it, they won’t do it.

The latest revision of the Council for the Accreditation of Educator Preparation Standards (2022) increased the focus on the InTASC standards and dispositions. Instead of simply stating the InTASC standards should be used, the current standards make this explicit in each of the CAEP standards. The first Standard (Content and Pedagogical Knowledge) identifies each of the 10 InTASC Standards individually, linking them to one of the four segments of Content and Pedagogical knowledge. Standard 2, Clinical Partnerships and Practices, requires experiences designed to develop knowledge, skills, and professional dispositions. The third standard, Candidate Recruitment, Progression, and Support, requires monitoring critical dispositions at transition points. The fourth standard, Program Impact, requires demonstration that program completers (post-graduation) apply dispositions in a P-12 classroom.

Previously the InTASC Standards were mentioned almost as in passing as a footnote or globally as the accreditation standards evolved. Now, interestingly, it is explicit and, even more intriguing, the CAEP standard calls for measures (in the plural). The literature, though, clearly points to single instrument use as the norm (LaPaglia, 2020). Next, we summarize the “state of the art” of single measures.

8.2.4 Representative unitary instruments in the literature

Institutions seeking teacher preparation accreditation tend to use single or unitary instruments, typically composed of selected response items. The literature provides many examples of single assessments of teacher affect such as surveys, indices, observations, or interviews (Richardson & Onwuegbuzie, 2003; Lund et al., 2007; Schulte et al., 2004; Wasicsko, 2004; Jung & Vogt, 2006; Singh & Stoloff, 2008). More recently, West et al. (2020) validated a *Teacher Disposition Scale*.

The Schulte index was stated to be based on InTASC Standards with content validity determined through expert panel review and factor analysis, but no systematic approach identified relative to the standards. The Singh and Stoloff (2008) index was built by "borrowing ideas from the existing indices of dispositions" (p. 1172) as well as from InTASC principles with validity and reliability data not reported. The West et al. (2020) scale was developed by an expert panel of Highly Accomplished Teachers who identified five dispositions: Motivation to Teach, Teacher Efficacy, Willingness to Learn, Conscientiousness, and Interpersonal and Communications – each with associated traits. Advanced measurement modeling is employed in only one of these instruments, and none are systematically aligned with the national standards of teaching proposed by the CCSSO through the InTASC Standards (2013).

Missing from the literature is any assessment process that combines multiple measures, standards-based construct definition, taxonomic use, and modern measurement theory. This chapter fills that void with the four elements applied in a nine-step process with teacher education as the exemplar.

8.3 The nine-step process to innovative affective measurement

The nine steps applied in the DAATS development and implementation process are listed and then described in the remainder of this chapter:

- Step 1: Define the Construct Based on Relevant Standards
- Step 2: Identify Appropriate Taxonomy to Frame Item Development and Scoring
- Step 3: Identify and Explore Potential Multiple Instruments in Varied Formats, for example,
  - Thurstone Scale
  - Constructed Response Prompts/Reflections
  - Projective Apperception Test
  - Observation Checklist
  - Focus Groups
- Step 4: Develop Multiple Instruments Using Standards-Based Items Anchored in the Selected Taxonomy

- Step 5: Combine Item Types for a Single Estimated Result
- Step 6: Control for Scoring Error
- Step 7: Collect Evidence of Validity and Reliability for Individual and Combined Tests
- Step 8: Use Instruments to Answer Emerging Research Questions while Contributing to the Validity Argument
- Step 9: Reconsider Item Analysis in Terms of Assessment Purpose: Quality Improvement and Assurance

8.3.1 Step 1: define the construct based on relevant standards

The InTASC Standards, previously named Principles, have evolved over the decades, and we previously asserted that the focus on standards-based teacher dispositions is now even more deeply embedded in the accreditation standards than ever before. We continue this discussion with the dispositions standards themselves.

The current 2013 version of the InTASC Standards (CCSSO, 2013) takes the requirements to the highest level ever. The initial Standards (Principle) targeted new teachers only, and the earlier name for the principles was INTASC (with a capital "N") reflecting the word "new" in the title. "New was drop," the 'n' was transitioned to lower case, and the expectations for beginning, middle, and advanced teachers were articulated in detail. The recent addition of the descriptor 'critical' to dispositions indicates their increasing importance in this complex and detailed set of standards.

There are a total of 10 InTASC Standards, and they are organized into four categories as follows:

Group One: The learner and learning
#1 Learning development
#2 Learning differences
#3 Learning environments
Group Two: Content
#4 Content knowledge
#5 Application of content
Group Three: Instructional practice
#6 Assessment
#7 Planning for instruction
#8 Instructional strategies
Group Four: Professional responsibility
#9 Professional learning and ethical practice
#10 Leadership and collaboration

To illustrate, examples of InTASC critical dispositions are drawn from Standard #1, Learner Development, at the new teacher level. There are four, with each focused on a non-cognitive behavior – “respects,” “is committed to,” “takes responsibility,” and “values”:

- 1(h) The teacher respects learners’ differing strengths and needs and is committed to using this information to further each learner’s development.
- 1(i) The teacher is committed to using learners’ strengths as a basis for growth and their misconceptions as opportunities for learning.
- 1(j) The teacher takes responsibility for promoting learners’ growth and development.
- 1(k) The teacher values the input and contributions of families, colleagues, and other professionals in understanding and supporting each learner’s development.

8.3.2 Step 2: identify an appropriate taxonomy to frame item development and scoring

The affective taxonomy (Krathwohl et al., 1964) was designed for framing instruction about beliefs, values, and attitudes. Thurstone (1928) defined attitudes to include a person’s inclinations, feelings, prejudice, bias, preconceived notions, ideas, fears, threats, and convictions about any topic. While the Taxonomy has not been applied typically to assessment development and analysis in systematic ways, it is proving to be a successful strategy for improving the variability in scaling affect, especially when used concurrently with scaling procedures developed by Thurstone (1928) and Rasch (1980). Bloom’s cognitive taxonomy has already been demonstrated as useful and consistent with construct-mapped rulers and student learning objectives where item content categories were established along taxonomic categories (Mohamed et al., 2008).

A variation of the taxonomy (Wilkerson & Lang, 2011), used in our research, classifies student affect into six levels. Since affective measurement differences from teaching affect, we have modified the original taxonomy by adding a “bottom” level to the Taxonomy – unaware, which is used to identify respondents who are at the “pre-receiving” level (Wilkerson & Lang, 2011). The difference between the original use, teaching, and our application, measurement, is that no teacher ever sets out to teach the absence of a construct, while measurement might, in fact, identify respondents who have not yet begun to acquire the belief.

To operationalize the taxonomic levels with the disposition instruments, we defined each one regarding the typical teaching behaviors that might be observed at each level including unaware. These are represented in Table 8.1.

For each assessment in the DAATS battery, candidates are classified on the levels of the Krathwohl Taxonomy, as we defined them. Overall, this translates to the following expectations for teachers and teacher candidates:

Table 8.1: Definition of taxonomic levels.

Taxonomic levels	Typical teaching behaviors at each taxonomic level
Unaware	– Has not considered the skill in any meaningful way. – May be opposed to the skill.
Receiving	– Recognizes (is aware of) importance. – Is beginning to think about it. – May provide a promise to use it without evidence of having used it.
Responding	– Is emotionally ready to do something and makes an attempt. – Gives a little extra effort, as time permits, to comply. – Can easily be distracted from application. – Has a beginning level of commitment or satisfaction.
Valuing	– Accepts worth and derives definite satisfaction from it. – Feels a need and would commit continuing time and effort. – Tolerates and may expect interferences.
Organization	– Plans, organizes, and schedules to ensure success with it. – Determines inter-relationships among knowledge and skills. – Adapts other aspects to fit it. – Is uncomfortable with interferences or lack of time to finish.
Characterization	– Sees the skill as the center or driving force of all work. – Helps others to see the skill’s importance, lobbying for it. – Integrates everything with it.

- Characterizing is interpreted as the level of mastery of a master teacher or leader and is assigned a rating of 5.
- Organizing is the target level for teacher leaders and is assigned a rating of 4.
- Valuing is the target level for beginning teachers and advanced teaching candidates and is assigned a rating of 3.
- Responding is the level of teacher candidates who have completed about half of their coursework and is assigned a rating of 2.
- Receiving is the level of teacher candidates at the time of admission to the program or after their first course and is assigned a rating of 1.
- Unaware is a rating of concern that requires attention and improvement and is assigned a rating of 0.

8.3.3 Step 3: identify and explore potential multiple measures in varied formats

8.3.3.1 The need for multiple measures

Herman et al. (2004) stated in clear terms that multiple measures are a necessity in cognitive assessment, and there is no reason to assume it is otherwise in the affective domain:

> No single test can tell all there is to know. As the directors of the National Center for Research on Evaluation, Standards, and Student Testing emphasize, "Multiple measures are needed to address the full depth and breadth of our expectations for student learning" (p. 2). Beyond the multiple-choice and short-answer items that are typical of current assessments, "other types of performance measures – essays, applied projects, portfolios, demonstrations, oral presentations, etc. – are needed to represent and guide students' progress." (Herman et al., 2004, p. 2)

There is no doubt that a single measure of cognition is not enough. The same holds true in the affective domain. One measure does not tell all, but the difference lies in the types of instruments that are appropriate for cognitive versus affective. We articulated this in our volume: the DAATS model (Wilkerson & Lang, 2007).

For single measures, especially if they are selected response like surveys, respondents might anticipate the "correct" answer (a socially acceptable response), whether or not they believe it (Edwards, 1959), so concerns about the easiest of the measures have brought fear into the room. However, this inherent deficit in the single measure (survey) suggests the need to take advantage of some potentially more revealing responses such as open-ended questions, observations, and interviews of impacted populations (Wilkerson & Lang, 2007; Lang & Wilkerson, 2008; Wilkerson & Lang, 2008).

Methods to measure the affective domain have clearly been available for many years (Thurstone, 1928); yet, despite substantial interest in measuring dispositions (e.g., diversity, multiculturalism, and social justice), there has been resistance based on measurement difficulties with particular resistance to moving toward multiple measures (Wilkerson & Lang, 2011).

Brindle (2012) recommended the use of varied methods to assess dispositions, providing students with ongoing feedback regarding dispositions, employing multiple assessors including student self-assessment, creating remediation plans when needed, and stressing the value and role of dispositions in effective teaching. We, too, have long advocated for multiple measures, noting that like knowledge and skills, we cannot rely on a single measure or a single point in time (Wilkerson & Lang, 2006; Lang et al., 2018a, 2018b, 2018c).

The combination of dichotomous and polytomous items on differing instruments can yield a single and useful score, or individual scores on each measure, that supports evaluation decisions that are useful, feasible, and accurate (Wilkerson, 2012). These bolder steps of adding constructed response items, however, require decisions and ratings that take time to develop with rubrics and trained raters. Then, implementation needs to be evaluated to determine if they can yield consistent results across raters. This also requires that measurement systems can incorporate or calibrate different item types representing the construct, and taxonomic scoring can be applied simultaneously.

We summarize our work in creating and combining multiple measures next, beginning with selected response methods. These are the most popular because they are relatively easy to create and score, taking the least time to administer. We will also provide relevant literature on each design and some limited early examples of items we developed and scored.

8.3.3.2 Selected response methods

Selected response methods provide an opportunity for respondents to respond based on a pre-determined set of responses item-by-item, like traditional testing in the cognitive domain. As with cognitive tests, guessing, or in this case, faking, the response is a deficit. In the affective domain, the respondent often agrees, rates importance, or makes some other value-laden judgment for pre-determined characteristics, for example, a belief in cultural diversity. Scales are an important method for measuring affect in this way (Anderson, 1988a). As with multiple choice items, these items can be relatively difficult to write well but easy to score.

There are four types of scales generally used: Thurstone agreement scales (Thurstone, 1928; Anderson,1988b), Likert scales (Anderson, 1988c), rating scales (Wolf, 1988), and semantic differential scales (Phillips, 1988). Thurstone scales are what we recommend, but here we provide descriptions of each of the four types, taken from our previous work. Table 8.2 provides descriptions and examples of the four scale types just identified, with Thurstone being the scale of our choice.

8.3.3.3 Constructed response methods

We include three constructed response methods: questionnaires, interviews, and focus groups, reproduced here as Table 8.3. Of course, only one (questionnaires) provides for written responses, so it is the most typical, being the easiest of the three to administer (Lang et al., 2019). Two require face-to-face interaction, so they are time-consuming and difficult to administer and score.

8.3.3.4 Observed performance

Observation assessment is another excellent source of data (Stalling & Mohlman, 1988). Included in this assessment classification are two types of observations of teachers in

Table 8.2: Selected response scales and examples.

Type of scale	Description	Example	
Thurstone Attitude	A set of statements are provided to respondents. They must agree or disagree with the statements, which typically number at least 20–45. Statements provide for a range in the attribute being measured.	1 = agree 0 = disagree	All children can learn. If I give it my best, most children can learn when given enough help.
Likert Scales	The same kinds of statements are provided to respondents as with Thurstone; however, in a Likert Scale, respondents indicate their agreement with statements on a five-point scale ranging from strongly agree to strongly disagree.	5 = Strongly agree 4 = agree 3 = neutral 2 = disagree 1 = strongly disagree	All children can learn. If I give it my best . . .
Rating Scales	These scales are very similar to Likert scales but allow for more flexibility in the response options, for example, a range from “like me” to “not like me” or “very important” to “not important.” Some rating scales only define end points, with only numbers in between anchors such as “dull” to “stimulating.”	4 = critically important 3 = very important 2 = somewhat important 1 = not important	All children should learn. I should give it my best, and if I do, most children will learn.
Semantic Differential Scales	These scales are similar to rating scales but often omit the numbers, leaving blanks instead between bipolar adjectives.	5 = critically important 1 = useless	I should give it my best, and if I do, most children will learn.

Source: Wilkerson and Lang (2007). *Dispositions Assessments Aligned with Teacher Standards*, Corwin Press, pp. 27–28.

the classroom. First is a planned and formal behavioral checklist (completed after multiple observations and products have been analyzed) and unplanned event reports (completed after observation of a specific targeted behavior); see Table 8.4.

In teacher education, it is standard to conduct observations of teacher skills in the classroom often throughout the program of studies. Observational techniques also provide an opportunity to look for evidence of specific behaviors that reflect the required or expected dispositions. The difficulty here, as in the other methods, is to ensure that affect and not skill is the construct evaluated. Observation is also venerable to judge errors.

Although not a formal part of the literature on affective measurement, one additional strategy that can work is to keep a record of events that occur that are triggers for concern about dispositions. This type of observation is of particular use in teacher assessment, and we call it the “disposition event report.” Using a taxonomy and scoring guide, these reports are still scorable as extreme items while recognizing they may not be calibrated with other measures.

Table 8.3: Descriptions and examples of constructed response methods.

Type	Description	Example
Questionnaires	Respondents are asked a pre-determined set of questions, the responses to which often require analysis, training, examples of responses, and rubrics. The questions are designed to elicit a high level of specificity to help ensure that the "correct" response is not given with a yes/no answer and to help sort out cheating and faking from genuine commitment. The assumption here is that if a teacher can describe the behavior in some detail, he/she is probably telling the truth and values the disposition enough to have it be a part of his/her work life.	Describe the last time you talked to a colleague about a problem with a student. What did the colleague recommend; what did you do; did it work; would you go back to that colleague again for advice? Note: If teachers don't value the advice of colleagues, they don't seek it.
Interviews	The same type of questions used in a questionnaire can be administered orally. The advantage is that the teacher can't check with others and fake a response as easily. The disadvantage, of course, is time. Interviews take longer.	Same questions and same assumption as for the questionnaire.
Focus groups	Questions are written for children to answer in a small group, and several can answer at the same time. These questions elicit data about how children perceive the teacher as a window onto the teacher's beliefs. Yes/no answers are acceptable in some instances.	Does your teacher listen to everyone in the class? Does he/she become impatient when you do not understand?

Source: Wilkerson and Lang (2007). *Dispositions Assessments Aligned with Teacher Standards*, Corwin Press, p. 29.

8.3.3.5 Projective techniques

Projective techniques (Walsh, 1988) are the last of the methods discussed that is useful for measuring dispositions, but they are also the most difficult of the strategies. Rorschach and Thematic Apperception Tests (TAT) are the most common examples of projective instruments used in psychology. A modified version of the original shown to respondents is shown in Table 8.5 as an example of the deliberate intention to evoke a reaction. The teacher is shown a picture that can evoke a variety of reactions about teaching and children, and the response indicates positive or negative values or beliefs, based on how the teacher "sees" the picture. Intense training and scoring procedures are required to interpret results.

Table 8.4: Self-report in DAATS.

Type	Description	Example
Observations	One or more performances can be evaluated in person in the classroom, looking for specific affective characteristics that are rated or counted.	Frequency count (tally) of teacher recognition of majority and minority children
Behavioral checklists	Items may be the same as the types on the observation instrument, but an overall impression of frequency is provided.	3 = frequently 2 = occasionally 1 = rarely Recognition of minority children
Event report	The event report is completed on an ad hoc basis when a person in authority (professor, mentor, principal) observes (or learns of) an action on the part of the teacher that is inappropriate. The report is used only for serious incidents, needs to be documented carefully, and should include follow-up between the author and the teacher to attempt to remediate the problem. Disciplinary action may be a part of the reporting system, depending on the gravity of the event. Disposition Event Reports could also be used to describe extremely positive events, but this is rare.	Racist remark, physical contact with a child, cheating, written reflection that indicates all students are dumb or certain students are incapable of learning, continuous late arrival for class, continuous inappropriate dress

Source: Wilkerson and Lang (2007). *Dispositions Assessments Aligned with Teacher Standards*, Corwin Press, p. 30.

Table 8.5: Apperception in DAATS.

Type	Description	Example
Teacher Apperception Test	Teachers are shown a picture and asked to say what they see. The interviewer records their reactions, searching for evidence of teaching values in their responses.	Teacher is shown a picture of a young woman dressed in a very short skirt with a low-cut blouse and body jewelry. The prompt is: "Tell me about this teacher." Most teachers would discuss the clothing, but if not, an additional prompt might be necessary. A teacher with appropriate values would question the attire; a teacher with inappropriate values would admire it.

Source: Wilkerson and Lang (2007). *Dispositions Assessments Aligned with Teacher Standards*, Corwin Press, p. 29.

The big idea here is that effective assessment techniques already exist for creating affective measures, and they provide opportunities for the use of a variety of item types that are increasingly difficult but revealing, providing for increasing levels of inference that increase the confidence decisions.

8.3.4 Step 4: develop multiple instruments using standards-based items anchored in a taxonomy

The DAATS battery (Wilkerson & Lang, 2006), creating using the process described in our book (Wilkerson & Lang, 2007), is composed of five standards-based instruments of different item types. Extensive discussion of the literature, the model, three of the instruments, and the results has been presented (Wilkerson & Lang, 2004; Lang, 2008; Lang & Wilkerson, 2008; and Englehart et al., 2012; Wilkerson & Lang, 2011; Wilkerson, 2012). The instruments are:

- Selected response (Thurstone Scale): *Beliefs About Teaching Scale*, version 2 (BATS2)
- Constructed response: *Experiential Teaching Questionnaire*, version 2 (ETQ2)
- Projective apperception test: *Situational Reflection Assessment*, version 2 (SRA2)
- Observed performance: *Candidate Behavior Checklist*, version 2 (CBC)
- Focus group: *K-12 Impact Disposition Scale,* (KIDS)

8.3.4.1 Beliefs About Teaching Scale, version 2 (BATS2)

BATS2 is a Thurstone (1928) dichotomous agree/disagree scale. Both Likert and Thurstone scales may be composed of statements to which respondents agree or disagree, but Thurstone's technique requires a dichotomous decision (agree/disagree only), while Likert provides for a rating scale, typically five points from strongly agree to strongly disagree with a neutral midpoint.

We selected the Thurstone approach based on support in the literature. Andrich (1988) concluded that the [Likert] approach is popular because of its simplicity. He wrote that it is simple, focusing directly on attitudes and that researchers find it satisfactory because it is "theoretically undemanding" (p. 12). However, Andrich concluded that he preferred Thurstone's approach, combining it with a probabilistic scaling model (Rasch, 1980).

Roberts et al. (1999), too, examined the relationship between Likert and Thurstone agreement scaling, recommending the Thurstone scale when extreme positions (e.g., high/low levels of commitment) are of interest:

> The Likert procedure may falter for individuals who hold extreme attitudinal positions when responses result from some type of ideal point process. This is because the Likert procedure is functionally a cumulative model of the response process, and as such, it is not always compatible with responses from an ideal point process. In contrast, the Thurstone procedure is functionally an unfolding model, and thus, it does correspond to the situation in which responses follow from an ideal point process. Due to this correspondence, the Thurstone procedure does not suffer

> from the degraded validity exhibited with the Likert method when individuals with extreme attitudes are measured. (pp. 229–230)

In our work, these dichotomous items are scored using the Winsteps software, based on whether we expect the candidate to agree or disagree, which we then interpret as consistent or inconsistent with the InTASC Standards.

BATS items are written with the intent of providing for a wide range of difficulty. There are, on average, six items per standard, and there are two versions of the test (Form A and Form B). The scales are calibrated using Winsteps (Linacre, 2023).

An example of two items, the taxonomic levels, and the expected responses for Standard #2 follows. The critical disposition is: *The teacher respects learners as individuals with differing personal and family backgrounds and various skills, abilities, perspectives, talents, and interests*:

> I usually think about children’s home life and environment so that I can tell if something is wrong. [Valuing; expected response is agree]
>
> I have a rule in my classroom: “We all speak proper English and ignore gestures, slang, or foreign languages.” [Unaware; expected response is disagree]

8.3.4.2 Experiential Teaching Questionnaire, version 2 (ETQ2)

The Experiential Teaching Questionnaire, version 2 (ETQ2) is a 10-item, guided-reflection that includes sub-sets of questions targeting the Critical Dispositions included in the InTASC Standards. There are currently two forms of the ETQ2 (Form A and B), and the questions on both forms directly align with each of the 10 InTASC Standards, and the associated Critical Dispositions. Items are designed to yield responses that span the affective taxonomy. It is scored manually and a little more difficult to fake than BATS2, so it provides the next level of inference and useful assessment of dispositions beyond the selected response Thurstone scale – the BATS2 self-report. The ETQ2 is more time-consuming to complete and to score than the selected response assessment (BATS2), but, using constructed responses, it provides the next level of useful assessment of dispositions and a clearer picture of what the teachers really believe about the Standards and their own behaviors related to them.

An example of the questions, guided by the Standards, is provided here for Standard 4 (Content Knowledge) reads: “The teacher understands the central concepts, tools of inquiry, and structures of the discipline(s) he or she teaches and creates learning experiences that make these aspects of the discipline accessible and meaningful for learners to assure mastery of the content.” For that Standard, Critical Disposition is “4(o) The teacher realizes that content knowledge is not a fixed body of facts but is complex, culturally situated, and ever evolving. S/he keeps abreast of new ideas and understandings in the field.” Associated ETQ2 reflective questions are:

- "How have you kept abreast of current developments in your field?
- Have you attended a workshop, staff development course, or conference that changed your teaching or views on teaching?
- Describe briefly what changed if anything."

The scoring guide for InTASC Standard 4 on Content Knowledge includes the exemplars provided in Table 8.6. The association of the sample responses with the continuum of ratings from 0 to 4 maps the expected variation in the construct. The validity of this scoring guide is evaluated for the extent to which each rating category defines a range in the measurement scale that makes the quantitative results qualitatively interpretable. When each rating category in turn becomes the most probable in the interactions of student performances and item difficulties, a key step toward construct validity is taken.

Table 8.6: Exemplars in ETQ Scoring Guide for Standard 4.

Rating	Sample response or exemplar
Unaware (0)	I expected the students to use their brains and think about the problem. They don't have a clue. I redirected and modeled making them questions in a critical thinking processing way of how I think.
Receiving (1)	I don't necessarily have students ponder or brainstorm things often, but lessons often do not go as I planned. This week alone I planned to spend 2 days on factoring and after 2 days the students were still having a lot of trouble, so I changed my plans for the rest of the week to allow more time. Once in a while I think, man, these kids aren't trying. But usually I realize that sometimes it takes students longer to learn something than I anticipated and that's okay.
Responding (2)	If a lesson was difficult or complex and they did not understand it, I would switch to something else and revisit it later. I would reflect on it and find another solution to the problem. Children come from different experiences and they may not share all the same experiences. Yes I would make changes and my interactions to help them "think." I would use more visuals, alter my language, and find more props to help them understand the concept.
Valuing (3)	When a plan doesn't go well it doesn't matter. Luckily I've been given a gift of creativity. Coming up with another way to teach the same information is usually easy for me. Changing to be a more effective teacher is good for the students. I taught one class after a snack and the children had energy. The first thing they did was an exercise which combines learning and running. When they were exhausted the second plan had them quietly sitting on the floor listening and resting.
Organizing (4)	This is a common occurrence that a writing teacher experiences. Writing is one of the highest forms of thinking, and most kids struggle with this intangible content area. Each year, I give a writing assignment that I plan but does not make sense to them. I strive for participation rewarding and praising highly those who execute the task. I also help them with graphic aids, vocabulary words, and even **create my own story** to share. I am constantly reworking my teaching to fit the particular class I have.

8.3.4.3 Situational Reflection Assessment, version 2 (SRA2)

The Situational Reflection Assessment, version 2 (SRA2) is composed of a series of 20 ambiguous sketches, designed by these authors but created by a professional artist. Unlike the traditional TAT, developed by Murray and Morgan in the 1930s (Murray, 1943), SRA2 provides not only the picture prompts but also some guiding questions designed to focus the teacher on specific InTASC Standards. In the early years, we attempted to use all 20, or at least 10 of the prompts, but there was resistance to the length of the test. At present, the pictures have been sorted into the four InTASC Categories, and alternate forms of the test are being calibrated.

In earlier versions of the test, each picture was associated with a single InTASC Standard. For example, the picture labeled "Walking to School" (Figure 8.1) included the prompt: "What kind of teacher would be best to teach this child? What would you do if this child were in your class?" It was used to provide a response to Standard 2, Learning Differences, to focus on diversity. However, the pictures are rich enough to be useful at the Category level. In this case: The learner and Learning, with Standards 1–3 – Learner and Learning (Standards 1–3. Learner Development, Learning Differences, and Learning Environments.) The picture and current prompt follow:

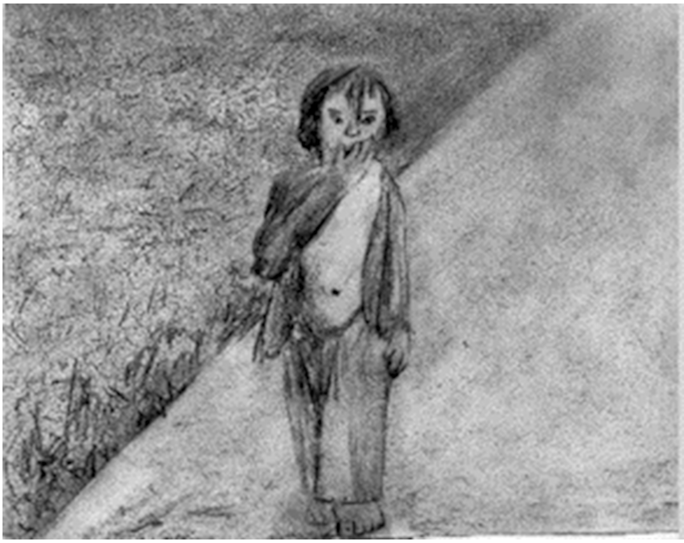


Figure 8.1: Sample SRA picture in DAATS. [Artwork by Barbara Slitkin.]

Think about the student in the picture who is different or struggling in some way.

- *Would you want this student in your class? Why or why not?*
- *In one or two sentences, tell the story of this student. Who are they?*
- *What pops into your mind as important to the success of this student?*
- *How do you feel when you see or think about this student? What is your first emotional reaction?*

Sample responses and scores mapping the expected variation in the measured construct for the "Walking to School" item are listed in Table 8.7.

Table 8.7: Sample responses scored on a Krathwohl taxonomy.

Rating	Sample response
Unaware (0)	This child seems to be from another country and has little time in the united states. An ESOL teacher would be best to teach this child because he may not know much English or much about the culture here and the best way for him to learn all of this would be to have a teacher whose focus is kids who don't know much of the English language
Receiving (1)	I can tell this child is of lower economic status due to the shirt and the absence of shoes. This child seems to have a lot of energy so a teacher with good physical tactics that can be incorporated into a lesson would fit best for him. If this student was in my class I would be very patient with him and try to see things from his point of view. I would talk to his parents about any problems he has I should know about and contact the principal to talk about concerns. I would allow this child in my class with great caution
Responding (2)	I can tell that this child does not come from a wealthy home. Their parents are not able to provide them with adequate clothing. A generous and accepting teacher (should be all) would be best to teach this child. If I was assigned a child like this in my class, I would contact the counselor's office and nurse's office to acquire adequate clothing for the student and to make sure they are being treated well at home. I want this child to stay in my class so I can make sure they get help and an education.
Valuing (3)	I can tell this child comes from poverty since he does not have a proper shirt or shoes. He does not have a book bag or any school supplies. A sincere teacher familiar with working with children that come from different backgrounds would be most effective in teaching this child. I would offer extra help to this child in both classwork and socially since he will most likely struggle with his classmates. I would talk to a guidance counselor about the right approaches to take with this child and see about the child meeting with the counselor too. I would want this child in my class. There is no child I would not want in my class. It would make my work as a teacher much more rewarding.
Organizing (4)	I can only assume that this student does not have the best home life. He has no shoes on and is either wearing a jacket and no shirt or a shirt that is simply hanging open. A nurturing teacher would be best for this child, one that wouldn't be afraid to monitor the situation and get involved on behalf of the child if needed. If this child was assigned to my class I would speak to the parents about what was going on and inform them that he has come to class barefoot and if nothing improves from there, I will take further action to get him the help that he needs. I would welcome any child with any home situation, behavioral issue, or disability to my classroom. They're all unique and we could help each other as the school year progresses. These students challenge you for the better and make for better teachers and they also need someone understanding, patient and ready to stand in the gap for them and make sure they get the help they need.

8.3.4.4 Candidate Behavior Checklist, version 2 (CBC2)

Like the other instruments in the battery, the Candidate Behavior Checklist (also called the Candidate Belief Checklist) draws items that are aligned with the InTASC standards. Early versions of CBC required a frequency rating of "typically positive," "mixed positive and negative," and "typically negative" for paired statements. For example, for Standard 8, on Assessment, one pair of items was:

> Encourages students with data-based or specific feedback.
> Makes disparaging remarks about individual or group generalized progress.

Virtually no students had a mixed or typically negative rating, and the linkages to the standards were determined to be too weak.

In 2021, the scale was re-written to align better with the Standards and with the Krathwohl Taxonomy. For example, for Standard 6, Assessment, disposition 6(t) states "The teacher is committed to using multiple types of assessment procedures to support, verify, and document learning." Item 13 on the CBC is "Builds formative and self-assessments, formal and informal, into all lessons."

A four-point scale was developed based on Krathwohl since experience with the other instruments indicated it was often difficult to differentiate between categories – much like the cognitive taxonomy that collapses each pair of levels, for example, knowledge and comprehension are often collapsed; application and analysis are often collapsed. For CBC, unaware remains alone, receiving/responding was collapsed, organizing and valuing were collapsed, and characterizing remains single but unexpected. The institution using CBC has not been forthcoming with producing the data since it is stored on a platform that does not permit aggregation, so no results are available.

8.3.4.5 K-12 Disposition Scale (KIDS)

KIDS, *K-12 Impact Dispositions Scale* (KIDS), uses a focus group technique, in which students participate in a focus group aimed at the dispositions of their teacher, combining both observation and rating. In other words, it assesses the teacher from the students perceptions but also based on the same construct (InTASC standards) as the self-report and observation instruments. KIDS is difficult to implement for a variety of reasons. Use of human subject requirements is rigid because of the interactions with children. Added to that are recordings and scoring requirements. However, when we have been able to use it, we have found that results provide rich, qualitative data that can be calibrated on the same scale as other items. As with the other instruments, the questions are written based on the InTASC Standards. For example, here is a sample disposition followed by sample question prompts:

> *The teacher appreciates individual variation within each area of development, shows respect for the diverse talents of all learners, and is committed to help them develop self-confidence and competence*:
> 1. What does your teacher do when you have trouble understanding?
> 2. How does your teacher get you interested in learning?
> 3. How do you know if your teacher likes to work with you?
> 4. Do you feel like you are learning in your class? Why? Why not?

Student responses can be scored in a similar way to the ETQ2 and SRA2 according to the Krathwohl taxonomy and also a qualitative validation of the score. For example, one teacher reported:

> The one question that the kids answered so cutely was if the teacher had any favorites. Almost all of them answered yes. I felt compelled to ask them who and it was themselves and their friends in the classroom. I thought the teacher was doing something right to make the kids feel like they all were her favorite.

Self-report items, questionnaires, and observations may conclude with the same estimate of the teacher dispositions, but sometimes the students provide a contrary view from other item types.

8.3.5 Step 5: combine item types for a single estimated result

The DAATS battery of instruments was investigated using a rating scale measurement model (Andrich, 1988) and Winsteps software, version 3.71 (Linacre, 2023). Items were combined into a single scale that included both dichotomous items (BATS) and rating scale items (ETQ and SRA). See Lang and Wilkerson (2008) for results of separate instrument calibrations and Wilkerson (2012) for combined scale calibration results.

Measurement modeling involves a careful delineation of the expected construct during the instrument design stage (Wilson, 2023). Conceptually, the idea of measurement modeling is simple. The ability (or, in this case, commitment) of individuals and the difficulty of items influence each other conjointly. The estimation process places them on the same interval scale, so predictions about the behavior of either one in relation to the other can be made. The model conceptualizes and estimates the mathematical relationship between a person's ability (or commitment) and the difficulty of an item, demonstrating that the probability of providing a correct response was related to the ability (or commitment) of the respondent.

Measurement modeling evaluates responses to dichotomously scored and rating scale items for the extent to which they support, within a fit-for-purpose uncertainty tolerance, the estimation of interval unit quantities. The availability of defined quantity estimates allows more appropriate use of common statistics, providing advantages over ordinal scores (counts) summed across correct responses. With a purposive sample and a skewed distribution, inferential statistics are not appropriate. Identified measurement models inform the mapping and estimation of constructs that experimentally tested for invariance across samples and that require neither large samples nor normal distributions (Bond et al., 2021). End users may then create interval level scales that can then be applied in associational or intervention research designs in subsequent studies. Validity and reliability statistics can also be reported (Linacre, 2023). Probabilistic measurement modeling has been extensively used in for decades in the development of major high-stakes tests (Kelley & Schumacher, 1984; Masters, 2007; Sabah et al., 2023; Samsudin et al., 2020).

8.3.6 Step 6: control for scoring error

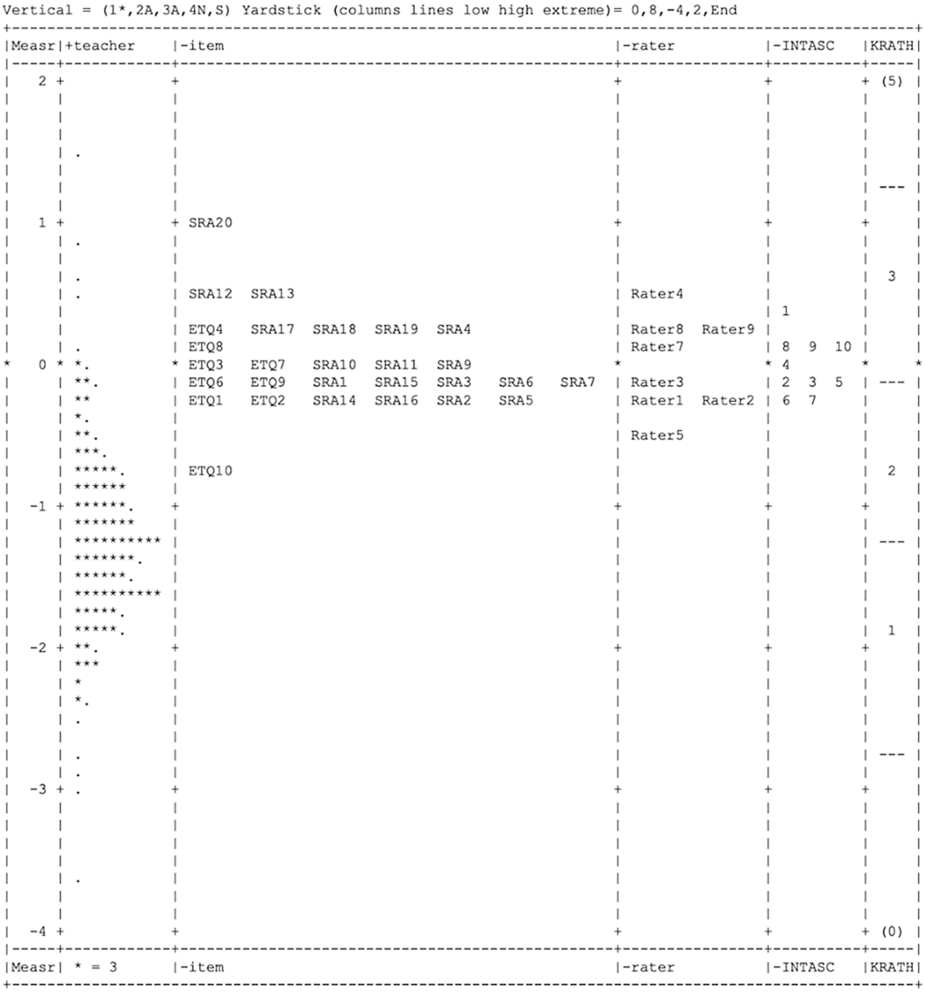
Rater effects pose a serious challenge to measurement (Engelhard, 1994; Wolfe, 2004) and can worsen or drift over time (Myford & Wolfe, 2009). Errors in rater judgment can impact the accuracy of ratings, and these effects are common but can be lessened through training of raters and monitoring of their efforts. Myford and Wolfe (2009) demonstrate how the nature of these effects can be modeled and understood. Given the clinical use of the DAATS battery, we needed to monitor rater performance as critical to valid assessment. Notably, we needed to combine judged items (like SRA2 and ETQ2) on the taxonomic scale used with multiple instruments. An analogy might be to assess verbal ability by judging an original short story along with asking the writer to compose an original piece on the spot, and scoring both activities together based on 10 standards, and utilizing Bloom's taxonomy. For this chapter, we will provide an example.

Figure 8.2 illustrates comparison of item types and judges rating a sample of beginning teacher candidates who took the SRA2 (apperception) and rated their reflections from a clinical experience (ETQ2). Both the SRA2 (20 items) and ETQ2 (10 items) were scored by nine raters on the same taxonomic scale, making the measurements based on them comparable despite their completely different item types. We used FACETS rulers to compare judges' leniency and looked to see if the SRA2 and ETQ2 were measuring the disposition construct (10 InTASC Standards) with two different item types (Lang et al., 2014, p. 246).

8.3.7 Step 7: present evidence of validity and reliability for individual and combined tests

8.3.7.1 Initial content validity in the design phase

The DAATS battery can be considered a high stakes instrument because the results are used for accreditation and sometimes intervention (Wilkerson & Lang, 2011). As such we were very concerned about validity from the start. In order to ensue alignment with the Standards, we wrote items for each standard and, with two other users



```
Vertical = (1*,2A,3A,4N,S) Yardstick (columns lines low high extreme)= 0,8,-4,2,End
+------------------------------------------------------------------------------------------------------+
|Measr|+teacher    |-item                                            |-rater          |-INTASC   |KRATH|
|-----+------------+-------------------------------------------------+----------------+----------+-----|
|   2 +            +                                                 +                +          + (5) |
|     |            |                                                 |                |          |     |
|     |            |                                                 |                |          |     |
|     |            |                                                 |                |          |     |
|     | .          |                                                 |                |          |     |
|     |            |                                                 |                |          |     |
|     |            |                                                 |                |          | --- |
|     |            |                                                 |                |          |     |
|   1 +            + SRA20                                           +                +          +     |
|     | .          |                                                 |                |          |     |
|     |            |                                                 |                |          |     |
|     | .          |                                                 |                |          |  3  |
|     | .          | SRA12  SRA13                                    | Rater4         |          |     |
|     |            |                                                 |                | 1        |     |
|     |            | ETQ4   SRA17  SRA18  SRA19  SRA4                | Rater8  Rater9 |          |     |
|     | .          | ETQ8                                            | Rater7         | 8  9  10 |     |
*   0 * *.         * ETQ3   ETQ7   SRA10  SRA11  SRA9                *                * 4        *     *
|     | **.        | ETQ6   ETQ9   SRA1   SRA15  SRA3   SRA6   SRA7  | Rater3         | 2  3  5  | --- |
|     | **         | ETQ1   ETQ2   SRA14  SRA16  SRA2   SRA5         | Rater1  Rater2 | 6  7     |     |
|     | *.         |                                                 |                |          |     |
|     | **.        |                                                 | Rater5         |          |     |
|     | ***.       |                                                 |                |          |     |
|     | *****.     | ETQ10                                           |                |          |  2  |
|     | ******     |                                                 |                |          |     |
|  -1 + ******.    +                                                 +                +          +     |
|     | *******    |                                                 |                |          |     |
|     | ********** |                                                 |                |          | --- |
|     | *******.   |                                                 |                |          |     |
|     | ******.    |                                                 |                |          |     |
|     | ********** |                                                 |                |          |     |
|     | *****.     |                                                 |                |          |     |
|     | *****.     |                                                 |                |          |  1  |
|  -2 + **.        +                                                 +                +          +     |
|     | ***        |                                                 |                |          |     |
|     | *          |                                                 |                |          |     |
|     | *.         |                                                 |                |          |     |
|     | .          |                                                 |                |          |     |
|     |            |                                                 |                |          |     |
|     | .          |                                                 |                |          | --- |
|     | .          |                                                 |                |          |     |
|  -3 + .          +                                                 +                +          +     |
|     |            |                                                 |                |          |     |
|     |            |                                                 |                |          |     |
|     |            |                                                 |                |          |     |
|     |            |                                                 |                |          |     |
|     | .          |                                                 |                |          |     |
|     |            |                                                 |                |          |     |
|     |            |                                                 |                |          |     |
|  -4 +            +                                                 +                +          + (0) |
|-----+------------+-------------------------------------------------+----------------+----------+-----|
|Measr| * = 3      |-item                                            |-rater          |-INTASC   |KRATH|
+------------------------------------------------------------------------------------------------------+
```

Figure 8.2: FACETS to compare item types and judges.

(an expert panel), validated the alignments. BATS items were also written to an expected taxonomic level.

For example, Standard 2, Learning Differences, Critical Disposition 2(m) states: “The teacher respects learners as individuals with differing personal and family backgrounds and various skills, abilities, perspectives, talents, and interests.” There are two BATS2 agreement items measuring this disposition:

- I usually think about children’s home life and environment so that I can tell if something is wrong. (Taxonomy level: valuing)
- I have a rule in my classroom: “We all speak proper English and ignore gestures, slang, or foreign languages.” (Taxonomy level: unaware)

In SRA2, candidates respond to one of four pictures accompanied by verbal prompts for this category of standards, and they are asked the following questions related to the support of learner differences including differing personal and family backgrounds:

- *Would you want this student in your class? Why or why not?*
- *In one or two sentences, tell the story of this student. Who are they?*
- *What pops into your mind as important to the success of this student?*
- *How do you feel when you see or think about this student? What is your first emotional reaction?*

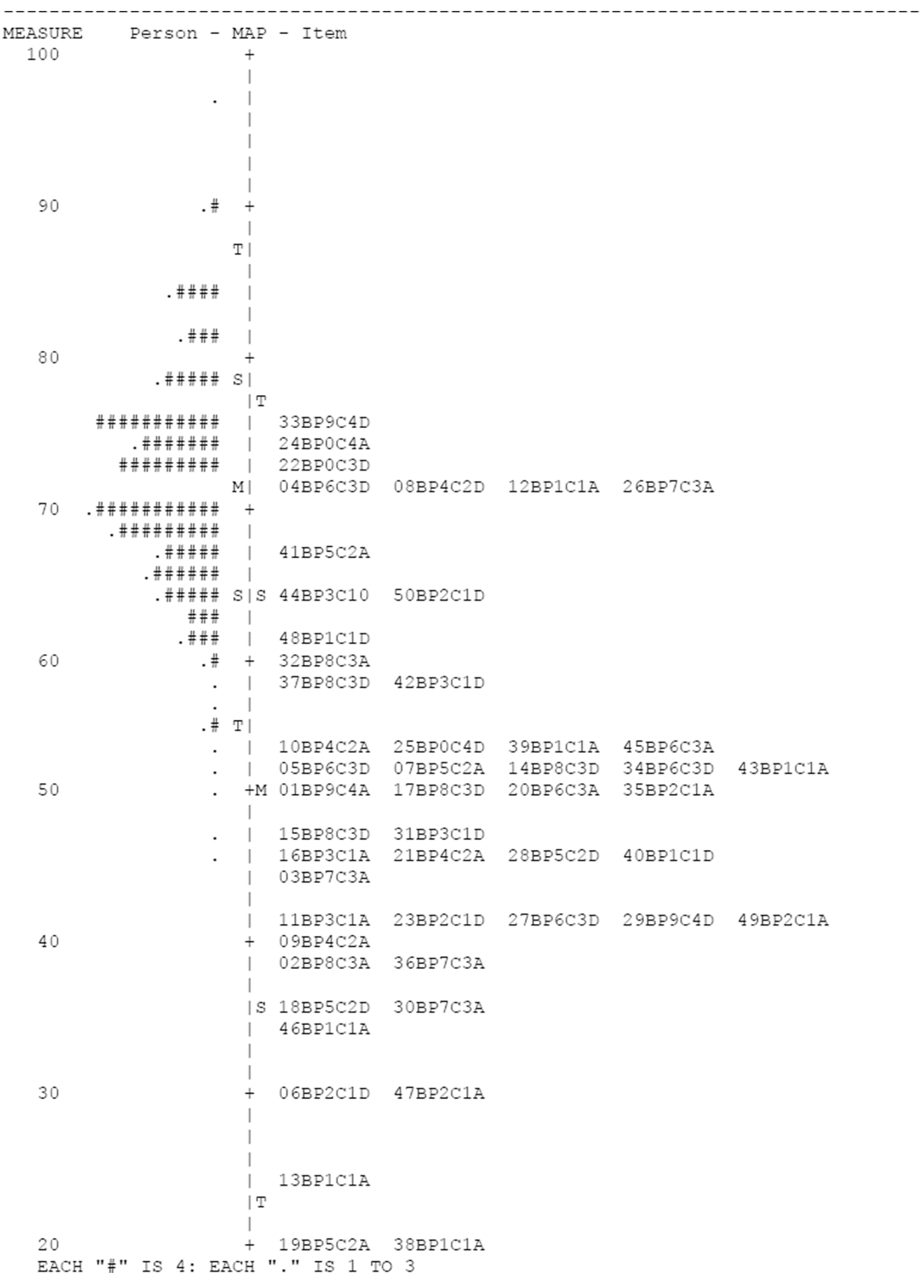
We continued our analyses of validity and reliability for both individual instruments and combined instruments over the years, often in combination with targeted research questions (Moore & Lang, 2023). We will return to those targeted research questions in Step 8. Here in Step 7, we present some selected basic measurement modeling techniques and results for individual and combined tests. The results we present represent an illustrative sampling of analyses we have done for individual studies and instruments. In each case, the source is provided for readers to explore further.

8.3.7.2 Graphic results through the Wright map

The variable (or Wright) map from Winsteps (Linacre, 2023) is often the starting point for interpreting a measurement analysis. A map presented for BAT2 (Form B), provided in Figure 8.3, shows the distribution of person commitment (left) and item difficulty (right). At the top are the most committed persons and most difficult items. The mean scale locations (M), and the locations of one (S) and two standard deviations (T) above and below the means, are shown for both persons and items.

The maps show normally distributed groups of persons and good coverage of the Standards. However, Figure 8.3 shows that the items are off target. The mean (M) of the item scale is more than two standard deviations below the mean of the student measurements. About half of the items are lower on the scale than 95% of the students. This occurs because there is limited representation of low scoring respondents since only teacher education candidates are in the pool. It has not been possible to test respondents with no interest in teaching. Items show good coverage of the construct in both forms.

The distributions of the items and persons, however, support confidence in the validity of inferences made from the measurements. In the following map, upper division students should find consistency with most BATS2 items easy after course work and internship experience. For example, an item such as “All students can learn” would have an expected response of agreement for students in teacher education programs.



```
--------------------------------------------------------------------------------
MEASURE    Person - MAP - Item
  100                +
                     |
                  .  |
                     |
                     |
                     |
                     |
   90            .#  +
                     |
                    T|
                     |
              .####  |
                     |
               .###  |
   80                +
             .##### S|
                     |T
        ###########  |  33BP9C4D
           .#######  |  24BP0C4A
          #########  |  22BP0C3D
                    M|  04BP6C3D  08BP4C2D  12BP1C1A  26BP7C3A
   70  .###########  +
         .#########  |
             .#####  |  41BP5C2A
            .######  |
             .##### S|S 44BP3C10  50BP2C1D
                ###  |
               .###  |  48BP1C1D
   60            .#  +  32BP8C3A
                  .  |  37BP8C3D  42BP3C1D
                  .  |
                 .# T|
                  .  |  10BP4C2A  25BP0C4D  39BP1C1A  45BP6C3A
                  .  |  05BP6C3D  07BP5C2A  14BP8C3D  34BP6C3D  43BP1C1A
   50             .  +M 01BP9C4A  17BP8C3D  20BP6C3A  35BP2C1A
                     |
                  .  |  15BP8C3D  31BP3C1D
                  .  |  16BP3C1A  21BP4C2A  28BP5C2D  40BP1C1D
                     |  03BP7C3A
                     |
                     |  11BP3C1A  23BP2C1D  27BP6C3D  29BP9C4D  49BP2C1A
   40                +  09BP4C2A
                     |  02BP8C3A  36BP7C3A
                     |
                     |S 18BP5C2D  30BP7C3A
                     |  46BP1C1A
                     |
                     |
   30                +  06BP2C1D  47BP2C1A
                     |
                     |
                     |
                     |  13BP1C1A
                     |T
                     |
   20                +  19BP5C2A  38BP1C1A
   EACH "#" IS 4: EACH "." IS 1 TO 3
```

Figure 8.3: Variable map of BATS2 Thurstone Scale (Lang et al., 2018a).

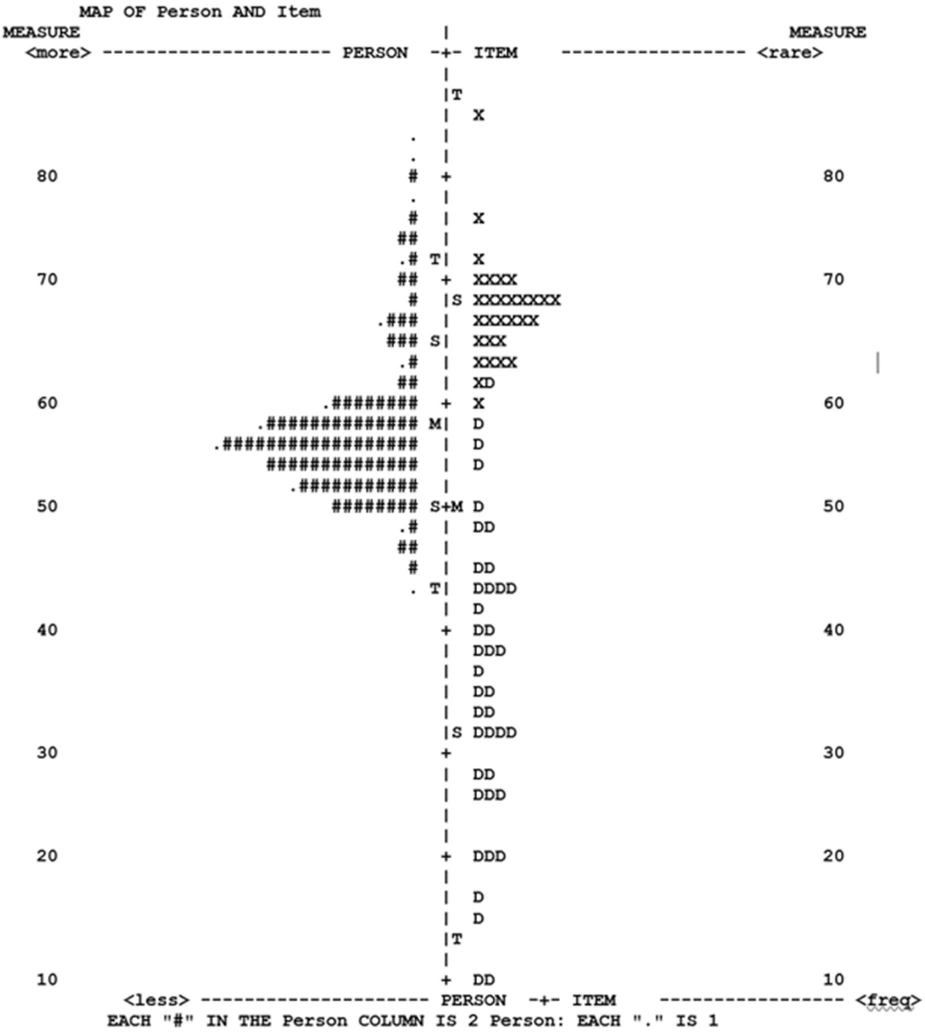
Item labels are coded so that the top item 33BP9C4D indicates that Item number = 33, Form = B, InTASC Principle (Standard) = 9, InTASC Category = 4, and Expected Response = D (disagree). Because the Standard (9) reflects Professional learning and ethical practice, but these students are preservice, they have minimal experience with this standard. At the other extreme, item 19BP5C2A is Principle 5 or Application of Content, which was a major part of student preservice training. The distribution of BATS2 with DAATS is often sample-dependent with the mean of calibrations changing from pre-internship undergraduates to upper division students to Inservice graduate students. This is related to training and experience.

BATS2 is generally used as a screening device (such as Form A on program entry and Form B on program exit) demonstrating improvement for accreditation. Some programs plan instruction based on group scores, and a few programs use scores for targeted internship placement. Item coding allows analysis by each of the 10 InTASC Standards or by the 4 InTASC categories.

We have had similar results for an analysis of combined instruments. The sample in this study included both undergraduate and graduate students. The Wright map for these items and respondents is shown in Figure 8.4. Here, again, the instrument's dichotomous items (indicated by the Ds in the right column) are off target toward the bottom of the scale. In contrast with Figure 8.3, though, some of the items (those with locations marked by *X*) have multiple rating categories. This means that the items at the *X* locations in Figure 8.4 are being shown only at their overall average calibration, though each of them is calibrated at every transition from one rating category to another. In this way, the rating scale augments the definition of the item hierarchy and supplements the interpretation of the measurements, in this case by substantively annotating performances lower and higher on the scale. In addition, Figure 8.4 shows three items that extend into the upper reaches of the student distribution, articulating their affective capacities at the high end of the standards, even for their mean values.

8.3.7.3 Descriptive, fit, and reliability statistics

Tables 8.8 and 8.9 provide descriptive, fit, and reliability statistics for items and people in the study of combined items. Fit is expressed in a mean square (MNSQ) statistic – the chi-square statistic divided by its degrees of freedom. The expected value should be close to 1.0. Values > 1.0 are considered underfit, introducing "noise," or another source of variance in the data. Values < 1.0 indicate over-prediction which can inflate other statistics. Infit is a *t*-standardized information-weighted mean square statistic, more sensitive to unexpected behavior affecting responses near the person's level; outfit is more sensitive to unexpected behavior on items far from the expected



```
       MAP OF Person AND Item
MEASURE                                 |                              MEASURE
  <more> --------------------- PERSON  -+- ITEM   ----------------- <rare>
                                        |
                                        |T
                                        |  X
                                     .  |
                                     .  |
   80                                #  +                                 80
                                     .  |
                                     #  |  X
                                    ##  |
                                    .# T|  X
   70                               ##  +  XXXX                           70
                                     #  |S XXXXXXXX
                                  .###  |  XXXXXX
                                   ### S|  XXX
                                    .#  |  XXXX
                                    ##  |  XD
   60                        .########  +  X                              60
                       .############## M|  D
                   .##################  |  D
                        ##############  |  D
                          .###########  |
   50                         ######## S+M D                              50
                                    .#  |  DD
                                    ##  |
                                     #  |  DD
                                     . T|  DDDD
                                        |  D
   40                                   +  DD                             40
                                        |  DDD
                                        |  D
                                        |  DD
                                        |  DD
                                        |S DDDD
   30                                   +                                 30
                                        |  DD
                                        |  DDD
                                        |
                                        |
   20                                   +  DDD                            20
                                        |
                                        |  D
                                        |  D
                                        |T
                                        |
   10                                   +  DD                             10
           <less> --------------------- PERSON  -+- ITEM    ----------------- <freq>
         EACH "#" IN THE Person COLUMN IS 2 Person: EACH "." IS 1
```

Figure 8.4: Variable map of BATS, ETQ, and SRA items combined (Wilkerson, 2012).

level. Fit in the range of 0.5–1.5 is considered productive. Fit > 2.0 degrades measurement; fit of 1.5–2.0 is unproductive, fit < 0.5 is overly predictable and potentially misleading. Z-Standardized (ZSTD) reports the statistical significance of the MNSQ and typically should not exceed 1.96 (Linacre, 2023).

Note that the mean for items was set to 50 and that the mean for persons is somewhat higher at 58. Ranges are strong: 11–84 for items 43–83 for persons. The standard deviation for items is almost two logits (18.4) and for people about one logit (10.9).

Table 8.8: Mean statistics for items.

	Total score	Count	Measure	Model error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
Mean	216.6	173.4	50	1.96	0.99	0	1.05	0.2
SD	80.6	18.4	18.22	1.31	0.13	1.4	0.27	1.6
Max.	358	186	83.82	7.11	1.43	3.5	2.05	3.5
Min.	77	85	10.82	0.9	0.64	−3.7	0.53	−3.7

Table 8.9: Mean statistics for persons.

	Total score	Count	Measure	Model error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
Mean	79.6	64	57.94	2.33	0.97	−0.2	1.03	0
SD	24.6	10.9	7.07	0.87	0.26	1.2	0.65	1.2
Max.	145	70	82.84	7.55	1.69	3	5.18	4
Min.	18	10	43.39	1.9	0.51	−2.7	0.23	−1.9

Mean fit statistics near the expected ranges of 1.0 for mean squares and 0.0 for standardized *z*'s are evident. Of the 70 items, only three exceeded the 1.5 outfit MNSQ (the highest was 2.05), and none exceeded 1.5 in infit. No items fell below 0.5 infit or outfit with the lowest at 0.64. These means, then, are not impacted by extreme scores. Just 11 respondents (6%) had an outfit MNSQ between 2.02 and 5.18. Modeling requirements are met.

Reliability and separation statistics are acceptable, given the uses of the instrument, with Cronbach's alpha (KR-20) estimated at 0.96, typically considered high because of missing data. This is consistent with the previous results. The person reliability of 0.87 indicates satisfactory separation of about three levels (separation = 2.67), which indicates that about four strata with centers three uncertainty ranges (standard errors) apart have been distinguished (for more on reliability, separation, and strata, see Linacre, 2013; Wright & Masters, 1982, pp. 92, 105–106; Wright & Masters, 2002). Item reliability and separation are good at 0.98 and 7.63.

8.3.7.4 Parallel forms and internal consistency

While we have analyzed BATS (versions 1 and 2) since its early development, we present here a more recent analysis in which we were testing parallel forms of the instrument (Lang et al., 2018a). We again used Winsteps software (Linacre, 2023), achieving the results presented in Table 8.10.

Person separation reliability (similar to internal consistency) indicates how well the scale discriminates between persons (Smith & Wind, 2018). The nature of the DAATS battery is such that some programs and institutions use short forms of one or

Table 8.10: Summary statistics, BATS2.

	Form A, *N* = 1072		**Form B, *N* = 372**	
	Separation reliability	**Outfit Z-Standard**	**Separation reliability**	**Outfit Z-Standard**
Persons	0.64 + α = 0.69	0.0 1.1 (SD)	0.66 + α = 0.69	0.1 0.8 (SD)
Items	0.99*	0.0 3.9 (SD)	0.97*	0.0 1.7 (SD)

+Indicates that the scale discriminates between persons adjusted for misfit in the data.
*Similar to internal consistency indicating items create a well-defined variable.
Source: Lang et al. (2018a).

two instruments with as few as four ETQ or SRA items for quick screening or class discussion. Following conventional recommendations (Smith & Wind, 2018), all instruments have been revised in their current forms to achieve expected person separation reliability with heterogeneous samples and trained judges (Lang, 2008; Lang et al., 2018a; Lang et al., 2021). Item separation has rarely been an issue in DAATS with values of 0.97–0.99 for most analyses, which is consistent with a well-defined variable and taxonomically aligned items. The overall precision obtained satisfies the needs of the decision process the tool is designed to support. Shorter forms implying lower reliability can meet the needs for screening or classroom diagnostics but would be inadequate to the demands of public defensible certification processes.

Note that the internal consistency reliability (for the items) of both instruments is excellent, and the ability of the instrument to distinguish people on the scale is adequate. In Table 8.10, the expected *Z*-Standard is 0.0 and the expected SD is 1.0. The only value that is not on target in the two forms is the SD of 3.0 in the Outfit *Z*-standard. This indicates that some items in the set may be misfitting. A subsequent review indicated four items that may need revision, but all four were substantially difficult for this population so the misfit may be partially due to the relatively extreme measure. A review and rewrite of misfitting items did not significantly alter the instrument pattern for extreme items, so they appear to be a consequence of position and a sensitivity to sample size. Dropping the four misfitting items and rerunning the analysis (Form A) did not change the measures of the top and bottom students more than 0.2 logits, and the Person Separation improved to 0.64 (from 0.55). The person separation is likely affected by the circumstances of the sample, where the length of the test is moderate, the sample ability is moderate, and the categories per item are restricted. These results conform with the pragmatic focus on the role of theory in motivating item composition and in evaluating empirical results advocated by Wright and Stone (1979) and adopted by others (Allen & Pak, 2023; Massof & Bradley, 2023).

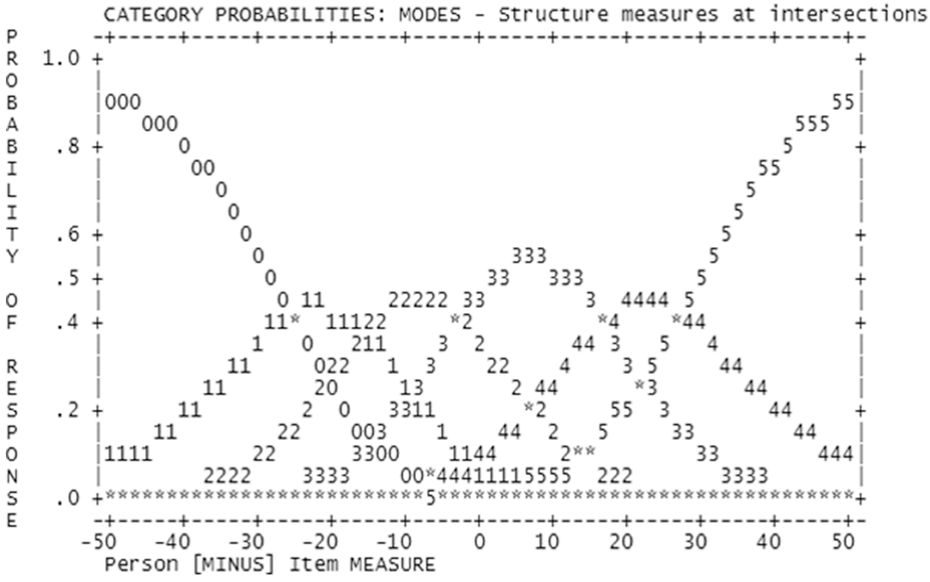
8.3.7.5 Ratings and category structure

The category structure of the rating scale items (ETQ and SRA) supports the use of ratings based on the affective taxonomy. The average measurements of respondents in each category increased in order, with fit statistics near the 1.0 target, as represented in Table 8.11.

Table 8.11: Category structure of ETQ and SRA.

Categ.	Count	%	Average	Infit MNSQ	Outfit MNSQ
0	446	9	−17.46	1.02	1.02
1	1,194	25	−13.52	0.96	0.95
2	1,765	37	−10.38	0.98	0.99
3	1,094	23	−4.15	0.95	0.97
4	190	4	4.74	0.83	0.85
5	25	1	10.95	0.89	0.9

Figure 8.5 represents graphically the category structure. Curves show how probable is the observation of each category for measures relative to the item measure and resembles the expected "range of hills" in which each category in turn becomes the expected rating for persons with measurements in the associated part of the continuum (Linacre, 2023).



```
        CATEGORY PROBABILITIES: MODES - Structure measures at intersections
P      -+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-
R  1.0 +                                                             +
O      |                                                             |
B      |000                                                        55|
A      |   000                                                  555  |
B   .8 +      0                                                5     +
I      |       00                                            55      |
L      |         0                                          5        |
I      |          0                                        5         |
T   .6 +           0                                      5          +
Y      |            0                    333             5           |
    .5 +             0                 33   333         5            +
O      |              0 11     22222 33        3  4444 5             |
F   .4 +             11*  11122     *2          *4    *44            +
       |            1   0   211    3  2       44 3   5   4           |
R      |          11     022   1  3    22    4    3 5     44         |
E      |        11       20     13       2 44      *3       44       |
S   .2 +      11        2  0   3311       *2     55  3        44     +
P      |    11        22    003    1    44  2   5     33        44   |
O      |1111        22      3300    1144     2**        33        444|
N      |        2222    3333    00*44411115555  222       3333       |
S   .0 +**************************5**********************************+
E      -+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-
      -50   -40   -30   -20   -10     0    10    20    30    40    50
        Person [MINUS] Item MEASURE
```

Figure 8.5: Threshold order of ETQ and SRA.

From the initial analyses until the current revisions and multiple forms, we consider evidence using standard guidelines (Bond et al., 2021; Smith & Wind, 2018) with the addition of an overarching taxonomy. Virtually every basic assessment textbook describes instrument planning using two-way blueprints or tables of specification (Popham, 2023) employing taxonomies to ensure content coverage, balance of item types, and construct modeling. A rule of DAATS construction was the use of a taxonomic plan for item development with our modified Krathwohl scale. We used category statistics in the manner of a rating scale model (Lang et al., 2014) so that calibrated scoring, regardless of item type, was aligned with the taxonomic plan. Response probability curves were an essential part of the DAATS battery development for both item construction and scoring guides.

8.3.8 Step 8: use instruments to answer emerging questions while contributing to the validity argument

We have not used the DAATS instruments in a vacuum – just to test a theory of how we can measure teacher dispositions. Each test was conducted to identify student needs, support accreditation demands for program improvement, or to answer a specific research question. In this step we discuss some of those research applications.

The use of DAATS instruments as affective assessment in teacher education implies use of the results for diagnosis and improvement3.8 of programs and individuals; however, unlike in cognitive assessment, results can tell a "story" about a candidate that may be very different from other candidates based on the results for individual items. To the extent that the story can be explained logically, there is evidence of construct validation. Engelhart et al. (2012) provide an example of program improvement we will now describe.

8.3.8.1 Example #1: program improvement

As a technique with accreditation at its core, we wanted to know if our students were equally predisposed to any of the InTASC Standards or, if instead, the faculty should attend more to any individual Standards. To that end, we analyzed the results of the multi-instrument results on the Standards as sub-scores. Table 8.12 presents descriptive statistics and reliability for each INTASC Standard in the combined analysis. Model reliability for each instrument ranges from 0.85 to 95.

Collaboration (#10) was the most challenging Standard and diverse learners (#3) easiest. Faculty confirm that candidates often prefer working alone, resisting teamwork, so this was an acknowledged area for improvement. They are taught consistently the importance of adapting for diversity, affirming the work of the faculty on this disposition. These two Standards were approximately one-half logit (and standard deviation) from the mean, evidence of construct validity. The results generally make

Table 8.12: DAATS statistics by InTASC Standard means.

Item count	Mean measure	S.E. mean	SD	Median	Model separation	Model reliability	Principle
70	49.17	2.33	19.37	50.78	5.98	0.97	**
7	50.53	6.93	16.97	49.88	7.64	0.98	1
7	48.73	6.39	15.65	37.8	8.06	0.98	2
7	45.25	10.52	25.77	58.42	3.49	0.92	3
7	47.34	9.24	22.63	56.52	6.58	0.98	4
7	49.02	6.56	16.08	51.68	7.04	0.98	5
7	50.93	6.56	16.07	45.03	8.53	0.99	6
7	50.78	6.49	15.9	60.29	7.82	0.98	7
7	46.32	8.22	20.14	39.6	7.52	0.98	8
7	48.37	9.16	22.43	53.65	7.56	0.98	9
7	54.42	7.08	17.35	49.88	9.66	0.99	10

sense. Reliability statistics are provided above. Differences in overall scores for respondents were not statistically significant between gender and ethnic categories.

8.3.8.2 Example #2: finding individual candidate improvement needs and/or celebrating candidate success

The correspondence between faculty perceptions of students and DAATS results supports construct validity (Englehart et al., 2012; Lang & Wilkerson, 2008). In one study we analyzed two cases – Candidates 18 and 22 – to illustrate this point. For Candidate 18, faculty judged the candidate to be enthusiastic about teaching and high in the cognitive domain. DAATS pinpointed specific, but limited, needs for improvement. Overall, her measurement was higher than Candidate 22's, whose interactions with faculty and her own students were not as effective as expected. The difference in the DAATS scores reflected that the different instruments varied in their indications of dispositions for the two candidates – a good example of the need for multiple measures:

Candidate 18:

- Standard on Content Knowledge, SRA: "I am always interested in learning more about what I do, and how I can be better. Since being a teacher impacts more than just me, I feel it is my responsibility to continue to grow."
- Standard on Communication, SRA: "Even though this child might be a handful, I would like to have him in my class. I feel that I could help him by being a stable, caring role model for him. I wouldn't want to see him in the hands of a teacher who would just feed off of him or have him get thrown out of school."

Candidate 22:

- In an ETQ question asking about an example of changes made to a critical thinking lesson plan that did not work well, she wrote:

  > The other day I did a reading lesson on cause and effect. I was asking the children to find the cause in multiple paragraphs and only half of the children were thinking. I told them to look back in the paragraph and find an answer because I wanted everyone to think and find an answer . . . I was getting aggrivated [sic] because many of the students werent even looking at the book . . . I didn't change any plans because they were just being lazy and not looking back . . . ."

- Also in ETQ, she responded "not applicable" to questions about planning lessons and units and using assessment data. She also indicated an unwillingness to let children make multiple attempts at success or to change plans that were not working.

Even though individual instruments provided evidence that was more consistent with faculty perceptions, the overall measures by InTASC Standard were consistent with faculty perceptions, as is clear in Figure 8.6.

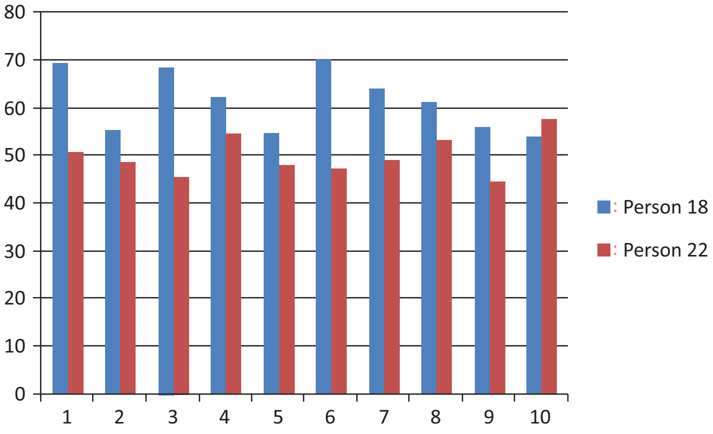


Figure 8.6: INTASC Standards measures for candidates 18 and 22.

The DAATS measures also serve as viable indicators of student acquisition of teacher dispositions during their program of study. LaPaglia and Wilkerson (2023) concluded that pre-admission dispositions can be improved as a function of cognitive coursework, but, since these improvements should not be assumed, they should be monitored for remediation purposes. Preservice teacher dispositions can change over time, based on instruction, and that those changes can be documented through the use of well-developed affective assessments.

8.3.8.3 Example #3: confirming cultural differences

During an exchange program, we collected BATS2 data from teacher education programs in two different cultures (Wilkerson et al., 2020). Using differential group functioning (DGF) reports from Winsteps (Linacre, 2023), we compared the dispositions between the groups by items and InTASC Standards (p. 119) and were able to explain group differences based on cultural norms, as hypothesized.

Table 8.13: Differential group functioning analysis by InTASC Standards.

Nation	DGF Size	Nation	DGF Size	DGF Contrast	Rasch-Welch t	DF	Probability	InTASC Standard
US	-.43	***China***	1.84	-2.27	-1.19	913	.2330	1
US	.00	China	-.76	.76	.46	675	.6433	2
US	-.51	***China***	3.11	-3.63	-3.13	979	.0018	3
US	-2.13	***China***	9.76	-11.89	-6.23	364	.000*	4
US	.32	China	-2.12	2.44	2.09	INF	.0370	5
US	.72	China	-4.99	5.71	4.66	925	.000*	6
US	-1.15	***China***	6.89	-8.04	-5.33	490	.000*	7
US	.27	China	-1.83	2.10	1.70	945	.0902	8
US	.20	China	-1.37	1.57	1.35	949	.1761	9
US	.32	China	-2.53	2.85	1.79	609	.0736	10

The results shown in Table 8.13 indicate that Standards 2, 5, 6, 8, 9, and 10 (Learning Differences, Planning for Instruction, Instructional Strategies, Application of Content, Professional Learning and Ethical Practice, and Leadership and Collaboration) were more difficult for Americans (positive values), while Standards 1, 3, 4, and 7 (Learner Development, Learning Environments, Assessment, and Content Knowledge) were more difficult (negative values) for the Chinese. Of these, only Standards 3, 4, and 7 (Chinese) and 6 (USA) were near or above the threshold set of one-half standard deviation (3.82 logits). Standards 4 (Content Knowledge) and 7 (Planning for Instruction) showing the most difficulty for Chinese compared to US candidates. Standard 6 (Assessment) indicated more difficulty for US candidates (Wilkerson et al., 2020).

8.3.9 Step 9: reconsider item analysis in terms of assessment purpose: quality improvement and assurance

In a reading or mathematics test, where the skills measured are readily identifiable and can be based on hierarchies of sub-skills, the measurement itself – or the score – is the goal. Whether Student A missed item 48 on a 50-item test or not may be of interest only if the incorrect response was highly unexpected. Such is not the case in the affective domain. Every item matter both individually and collectively because individual item responses could identify a disposition that might be harmful. That is a major issue in how we approach the use of a dispositions assessment, even in our approach with its multiple measures. We ask in this step of our model, what does the assessor do if the respondent writes something frightening? Hopefully, this will not happen, but we should be prepared, in any case.

If the purpose of the test is to diagnose and remediate at both the individual and program levels, as is the case with the DAATS instruments, then institutions engaging in this or a similar process for disciplines that require a combination of competence and caring, then it is important to plan for a more labor-intensive analysis of the results at the item level. As a whole, the DAATS instruments can validate the overall quality of institutional level instruction, providing the level of assurance needed for accreditation, but quality improvement could be at the sub-construct or even item level.

We offer one final example of our time working with an institution that used BATS. We have an item that states: “Every child can learn.” We placed this item at the receiving level on the Krathwohl Taxonomy since that fundamental principle – that belief in all children – is so deeply engrained in the teacher education culture, that we could not imagine a scenario in which a group of students would “disagree” with the statement. But it happened, and 68% of the tested candidates did, in fact, disagree. When the shock waves diminished, the administration and faculty embarked on a massive restructuring effort. If 68% of the students answered a multiplication item incorrectly but answered the remaining multiplication items relative well, the path would have been different. Maybe we would have looked for a bad distractor, but it would not have been earth shattering and the fix would have been much easier.

In dispositions assessment, every respondent, and every item matters.

8.4 Summary and conclusions

Beginning with an accreditation mandate to demonstrate compliance with the teacher education standards (quality assurance) while concurrently identifying areas needing strengthening (quality improvement), we developed an innovative approach to measuring teacher dispositions – a construct that has been evasive over the decades and has been the subject of hot debate.

We have committed to a standards-based approach, rooted in the professional standards of teaching over “habits of mind” and morality, developing a battery of assessments. We named the battery *Dispositions Aligned with Teacher Standards (DAATS),* after a book we wrote bearing the same title. The battery is composed of five instruments of different item types because the literature is clear on the need for multiple measures. Some instruments have been used more than others, and some have been used singly or in combination with others. We have answered research questions that intrigued us, while identifying strengths and weaknesses of individual candidates and programs in general.

In this chapter we have presented a nine-step process that incorporated four innovative ideas: (1) items should be written to show consistency with professional standards; (2) multiple instruments of different item types are necessary; (3) that design and scoring must be tied to a meaningful taxonomy that defines levels in the scale; and (4) modern measurement practice provides a solid foundation for use of the results in scoring candidates while seeking to improve both their individual performance as well as the performance of the program.

We have collected enough evidence to demonstrate the viability of scientific approaches, i.e., probabilistic models for measurement, to measure affect using instrument types previously endorsed in the literature. These ideas, operationalized through the nine-step process, have yielded positive results for us over the past 15 years and can provide an effective beginning to measurement professionals in both teacher education and in other disciplines that aspire to graduate students who are both competent and caring.

8.5 Limitations

Teacher education faculty have shown less commitment to dispositions than to competencies. It has been difficult to convince faculty that it is worth their time to read a constructed response to a dispositions-based prompt. Sampling, too, has been problematic; the scale (BATS2) is not so difficult to "sell," but ETQ and SRA require commitment on the part of faculty. It takes them too much time to learn to score and then score.

Sampling, too, is a limitation. We cannot measure the bottom of the scale. The candidates we measure have already selected teaching as their future career, so measuring candidates with no positive dispositions toward teaching is not possible. The colleges of business and engineering do not freely offer up their students to take our tests.

8.6 Recommendations

Without support from the Council for Accreditation of Educator Preparation (CAEP), the national accreditation agency for teacher education, the commitment of education faculty to assessing teacher dispositions in a meaningful way, is not likely to develop. While CAEP has taken great strides by integrating the InTASC Standards into their own, there is still no evidence that they really mean it at the knowledge, performance, and, most importantly, the critical dispositions level. So, recommendation #1 is that CAEP, as a body of teacher educators, consider whether competence and caring (knowledge/performance and critical dispositions) really are both important.

Beyond teacher education, our second recommendation is that other disciplines, especially those in the medical and human services fields, also consider whether they are committed to measuring "caring" and "commitment" in addition to "competence." If they are, then the techniques described herein should be tested in new ways and in new fields of study, with explicit recognition of the potential for extending the SI and the value that could be obtained from investments in metrology (Ashworth, 2004; Bernstein, 2004).

References

Allen, D. D., & Pak, S. (2023). Improving clinical practice with person-centered outcome measurement. In W. P. Fisher Jr. & S. J. Cano (Eds.). *Person centered outcome metrology* (pp. 53–105). Springer.

Anderson, L. W. (1988a). Attitudes and their measurement. In J. P. Keeves (Ed.). *Educational research, methodology, and measurement: An international handbook* (pp. 421–426). Oxford, England: Pergamon.

Anderson, L. W. (1988b). Likert scales. In J. P. Keeves (Ed.). *Educational research, methodology, and measurement: An international handbook* (pp. 427–428). Oxford, England: Pergamon.

Anderson, L. W. (1988c). Guttman scales. In J. P. Keeves (Ed.). *Educational research, methodology, and measurement: An international handbook* (pp. 428–430). Oxford, England: Pergamon.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Ashworth, W. J. (2004), Metrology and the state: Science, revenue, and commerce. *Science*, *306*(5700), 1314–1317.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions*, *22*(1), 1145–1146. https://www.rasch.org/rmt/rmt221a.htm

Bernstein, W. J. (2004). *The birth of plenty: How the prosperity of the modern world was created*. McGraw-Hill.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives, the classification of educational goals – Handbook I: Cognitive domain*. New York: McKay.

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. 2nd ed*.* Mahwah, NJ: Lawrence Erlbaum Associates.

Bond, T., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* 4th ed. NY: Routledge.

Borko, H., Liston, D., & Whitcomb, J. A. (2007). Apples and fishes: The debate over dispositions in teacher education. *Journal of Teacher Education*, *58*(5), 359–364.

Bradley, E., Isaac, P., & King, J. (2020). Assessment of pre-service teacher dispositions. *Excelsior, Leadership in Teaching and Learning*, *13*(1), 50–62.

Brindle, S. (2012). *An exploratory study on the assessment of pre-service teacher dispositions by teacher education programs in Iowa* [Doctoral dissertation, Drake University, Iowa].

Choi, H., Benson, N. F., & Shudak, N. J. (2016). Assessment of teacher candidate dispositions: Evidence of reliability and validity. *Teacher Education Quarterly*, *43*(2), 71–89. EJ1110316.

Council for the Accreditation of Educator Preparation. (2022). CAEP standards. CAEP. Downloaded from https://caepnet.org

Council of Chief State School Officers. (2013, April). Interstate teacher assessment and support consortium. In *InTASC model core teaching standards and learning progressions for teachers 1.0: A resource for ongoing teacher development*. Washington, DC: Author.

Dewey, J. (1944). *Democracy and education: An introduction to the philosophy of education*. New York: The Free Press.

Diez, M. E. (2007). Looking back and moving forward: Three tensions in the teacher dispositions discourse. *Journal of Teacher Education*, *58*(5), 388–396.

Dottin, E. (2009). A Deweyan approach to the development of moral dispositions in professional teacher education communities. In H. Sockett (Ed.). *Teacher dispositions: Building a teacher education framework of moral standards* (pp. 27–47). Washington, DC: AACTE.

Edwards, A. L. (1959). Social desirability and the description of others. *Journal of Abnormal and Social Psychology*, *59*, 434–436.

Englehart, D., Batchelder, H., Jennings, K., Wilkerson, J., & Lang, W. S. (2012). Teacher dispositions: Moving from assessment to improvement. *The International Journal of Educational and Psychological Assessment*, *9*(2), 26–44.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*, 93–112.

Herman, J. L., Baker, E. L., & Linn, R. L. (2004). Accountability systems in support of student learning: Moving to the next generation. In *CRESST line – Newsletter of the National Center for Research on Evaluation, Standards, and Student Testing*. Los Angeles, Calif: UCLA Center for the Study of Evaluation.

Jung, E., Rhodes, D., & Vogt, W. P. (2006). Disposition assessment in teacher education: A framework and instrument for assessing technology disposition. *The Teacher Educator*, *41*(4), 207–233.

Katz, L. G., & Raths, J. D. (1986). Dispositional goals for teacher education: Problems of identification and assessment. In *Conference papers* (pp. 1–29). Kingston, Jamaica.

Kelley, P. R., & Schumacher, C. F. (1984). The Rasch model: Its use by the National Board of Medical Examiners. *Evaluation & the Health Professions*, *7*(4), 443–454.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives: The classification of educational goals, Handbook II: Affective domain*. New York: David Mckay Company Incorporated.

Lang, W. S. (2008). The DAATS model: Initial psychometric and statistical findings. A top ten illustration. *Resources in Education*. (ERIC Document Reproduction Service No. ED502864).

Lang, W. S., LaPaglia, K., Moore, L., & Wilkerson, J. R. (2020, February). *Assessment of teacher dispositions meaningfully and credibly*. Paper delivered to the American Association of Colleges of Teacher Education, Atlanta, GA.

Lang, W. S., Moore, L., Wilkerson, J. R., Parfitt, C. M., Greene, J., Kratt, D., & Fields, L. (2018a). Beliefs about Teaching (BATS2) – Construction and validation of an instrument based on InTASC critical dispositions. *International Journal of Learning, Teaching and Educational Research*, *17*(8), 56–77. doi:10.26803/ijlter.17.8.4

Lang, W. S., & Wilkerson, J. (2008). Measuring teacher dispositions with different item structures: An application of the Rasch model to a complex accreditation requirement. *Resources in Education*. (ERIC Document Reproduction Service No. ED502965).

Lang, W., Wilkerson, J., Gilbert, S., Wang, C., Kratt, D., Parfitt, C., Martelli, D., Zhang, J., Fields, L., LaPaglia, K., Greene, J., & Johnston, V. (2018c). *Beliefs About Teaching (BATS2) – Construction and Validation of an Instrument Based on InTASC Critical Dispositions.* Paper presented at the annual meeting of the Eastern Educational Research Association, Clearwater Florida, February 8, 2018.

Lang, W. S., Wilkerson, J. R., & Moore, L. (2019). Assessment of teacher dispositions with the ETQ2: A guided-reflection and Rasch Model analysis. *Advances in Global Education and Research*, *3*(29), 301–308. ISBN 978-1-7321275-4-8.

Lang, W. S., Wilkerson, J. R., & Moore, L. (2021). *Beliefs about teaching (BATS2): Rasch model analysis of an instrument based on InTASC critical dispositions*. Paper delivered at the Global Conference on Education and Research, Tampa, FL.

Lang, W., Wilkerson, J., Moore, L., & Fields, L. (2018b). *Beliefs About Teaching (BATS2) – Assessment of InTASC Dispositions*. Paper delivered at the annual meeting of the American Association of Colleges of Teacher Education. Baltimore, MD. Paper delivered at the annual meeting of the American Association of Colleges of Teacher Education. Baltimore, MD.

Lang, W. S., Wilkerson, J. R., Rea, D. C., Quinn, D., Batchelder, H. L., Englehart, D. S., & Jennings, K. J. (2014). Measuring teacher dispositions using the DAATS battery: A multifaceted Rasch analysis of rater effect. *Journal of Applied Measurement*, *15*(3), 240–251.

LaPaglia, K. (2020). *Teacher dispositions: A case study exploring dispositional growth in a teacher educational program*. (Doctoral dissertation, Florida Gulf Coast University).

LaPaglia, K., & Wilkerson, J. R. (2023). Changes in preservice teacher dispositions during a teacher preparation program. *International Journal of Evaluation and Research in Education*, *12*(2), 1035–1040. ISSN: 2252-8822, doi:10.11591/ijere.v12i2.24224

Lei, P. W., Zhao, H., Hart, S. C., Li, X., & DiPerna, J. C. (2023). Examination of psychometric evidence for criterion-referenced scores from the SSIS SEL Brief Scales. *Journal of Psychoeducational Assessment*, *41*(3), 311–327.

Linacre, J. M. (1995). Paired comparisons with ties: Bradley-Terry and Rasch. *Rasch Measurement Transactions*, *9*(2), 425. http://www.rasch.org/rmt/rmt92d.htm

Linacre, J. M. (2000). Almost the Zermelo model? *Rasch Measurement Transactions*, *14*(2), 754. http://www.rasch.org/rmt/rmt142k.htm

Linacre, J. M. (2013). Reliability, separation, and strata: Percentage of sample in each level. *Rasch Measurement Transactions*, *26*(4), 1399. http://www.rasch.org/rmt/rmt264.pdf

Linacre, J. M. (2023). *A user's guide to WINSTEPS: Rasch-model computer programs, version 5.5.1*. Winsteps.com.

Lund, J., Wayda, V., Woodard, R., & Buck, M. (2007). Professional dispositions: What are we teaching prospective physical education teachers? *The Physical Educator*, *12*, 38–47.

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement*, *51*, 315–327. http://www.sciencedirect.com/science/article/pii/S0263224114000645

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement*. 2nd ed. Springer Series in Measurement Science and Technology. Springer. https://link.springer.com/book/10.1007/978-3-031-22448-5

Massof, R. W., & Bradley, C. (2023). An adaptive strategy for measuring patient-reported outcomes. In W. P. Fisher Jr. & S. J. Cano (Eds.). *Person-centered outcome metrology: Principles and applications for high stakes decision making* (pp. 107–150). Springer.

Masters, G. N. (2007). Special issue: Programme for International Student Assessment (PISA). *Journal of Applied Measurement*, *8*(3), 235–335.

Mohamed, A., Abdaziz, A., Zakaria, S., & Masodi, M. (2008, February). *Appraisal of course learning outcomes using Rasch Measurement: A case study in information technology education*. Conference: Proceedings of the 7th WSEAS International Conference on Software Engineering, Parallel and Distributed Systems https://www.researchgate.net/publication/262207688_Appraisal_of_course_learning_outcomes_using_rasch_measurement_a_case_study_in_information_technology_education

Moore, L. L., & Lang, W. S. (2023). Assessment of math teachers' dispositions to improve urban teacher-leaders' growth and effectiveness. *International Journal of Learning, Teaching and Educational Research*, *22*(4), 494–511. doi:10.26803/ijlter.22.4.27

Murray, H. A. (1943). *Thematic apperception test manual*. Cambridge, MA: Harvard University Press.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*, 371–389.

NCATE. (1987). *NCATE standards, procedures, and policies for the accreditation of professional education units*. Washington, D.C.: National Council for Accreditation of Teacher Education.

Newby, V. A., Conner, G. R., Grant, C. P., & Bunderson, C. V. (2009). The Rasch model and additive conjoint measurement. *Journal of Applied Measurement*, *10*(4), 348–354.

Niu, C., Everson, K., Dietrich, S., & Zippay, C. (2017). Validity issues in assessing dispositions: The confirmatory factor analysis of a teacher disposition form. *Southern Regional Association of Teacher Education*, *26*(2), 41–49. EJ1152459 http://www.srate.org/journal.html

Pancorbo, G., Primi, R., John, O. P., Santos, D., & De Fruyt, F. (2021). Formative assessment of social-emotional skills using rubrics: A review of knowns and unknowns. *Frontiers in Education*, *6*, 687661.

Pendrill, L. R. (2019). *Quality assured measurement: Unification across social and physical sciences*. Springer Series in Measurement Science and Technology. Springer. https://link.springer.com/book/10.1007/978-3-030-28695-8

Pendrill, L., & Fisher, W. P., Jr. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement*, *71*, 46–55. http://dx.doi.org/10.1016/j.measurement.2015.04.010

Phelan, A. (2001). The death of a child and the birth of practical wisdom. *Studies in Philosophy and Education*, *20*(1), 41–45.

Phelan, A. (2005). *On discernment: The wisdom of practice and the practice of wisdom in teacher education design: Developing a multi-linked conceptual framework* (pp. 57–73). Dordrecht: Springer.

Phelps, P. (2006). The dilemma of dispositions. *The Clearing House*, *79*(4), 174–178. https://www.jstor.org/stable/30181066

Phillips, J. L. (1988). Semantic differential. In J. P. Keeves (Ed.). *Educational research, methodology, and measurement: An international handbook* (pp. 430–432). Oxford, England: Pergamon.

Popham, W.J. (2023). *Classroom assessment: What teachers need to know*, 10th ed. New Jersey: Pearson Education.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Expanded ed. Chicago: University of Chicago Press.

Richardson, D., & Onwuegbuzie, A. (2003). Attitudes toward dispositions related to the teaching of pre-service teachers, in-service teachers, administrators, and college/university professors. (ERIC Document Reproduction Service ED482689).

Rike, C. J., & Sharp, L. K. (2008). Assessing pre-service teachers' dispositions: A critical dimension of professional preparation. *Childhood Education*, *84*(3), 150–153.

Roberts, J. S., Laughlin, J. E., & Wedel, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement*, *59*, 211–233.

Sabah, S., Akour, M. M., & Hammouri, H. (2023). Implementing next generation science practices in classrooms: Findings from TIMSS 2019. *Journal of Turkish Science Education*, *20*(2), 309–319.

Samsudin, M. A., Chut, T. S., Ismail, M. E., & Ahmad, N. J. (2020). A calibrated item bank for computerized adaptive testing in measuring science TIMSS Performance. *Eurasia Journal of Mathematics, Science and Technology Education*, *16*(7), em1863.

Schulte, L., Edick, N., Edwards, S., & Mackiel, D. (2004). The development and validation of the *Teacher Dispositions Index*. *Essays in Education*, *23*, 85–100.

Singh, D., & Stoloff, D. (2008). Assessment of Teacher Dispositions. *College Student Journal*, *42*(4), 1169–1180.

Smith, R. M., Julian, E., Lunz, M., Stahl, J., Schulz, M., & Wright, B. D. (1994). Applications of conjoint measurement in admission and professional certification programs. *International Journal of Educational Research*, *21*(6), 653–664.

Smith, R., & Wind, S. (2018). *Rasch measurement models: Interpreting WINSTEPS and FACETS output (2nd)*. Maple Grove, MN: JAM Press.

Stalling, J. A., & Mohlman, G. G. (1988). Classroom observation techniques. In J. P. Keeves (Ed.). *Educational research, methodology, and measurement: An international handbook* (pp. 469–173). Oxford, England: Pergamon.

Stenner, A., & Horabin, I. (1992). Three stages of construct definition. *Rasch Measurement Transactions*, *6*(3), 229.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529–554.

Walsh, J. J. (1988). Projective testing techniques. In J. P. Keeves (Ed.). *Educational research, methodology, and measurement: An international handbook* (pp. 432–436). Oxford, England: Pergamon.

Wasicsko, M. M. (2004). The 20-minute hiring assessment. The School Administrator Web Edition. [on-line site] http://aasa.org/publications/sa/2004_10/wasicsko.htm

West, C., Baker, A., Ehrich, J., Woodcock, S., Bokosmaty, S., Howard, S., & Eady, M. (2020). Teacher Disposition Scale (TDS): Construction and psychometric validation. *Journal of Further and Higher Education*, *44*(2), 185–200. doi:10.1080/0309877X.2018.1527022

Wilkerson, J. R. (1987). *SIDPASS, an accreditation self-study model with an initial application to National Council for Accreditation of Teacher Education (NCATE)*. Doctoral Dissertation University of South Florida.

Wilkerson, J. R. (2006). Measuring teacher dispositions: Standards-based or morality-based? *Teachers' College Record*, *20*, 85–95.

Wilkerson, J. R. (2012). Measurement and evaluation perspectives on scaling teacher affect with multiple measures. *The International Journal of Educational and Psychological Assessment*, *9*(2), 165–187.

Wilkerson, J. (2019). Rubrics meeting quality assurance and improvement needs in the accreditation context. *Quality Assurance in Education*, *28*(1), 19–32. https://doi.org/10.1108/QAE-04-2019-0045

Wilkerson, J. R., & Lang, W. S. (2004). A standards-driven, task-based assessment approach for teacher credentialing with potential for college accreditation. *Practical Assessment, Research, and Evaluation*, *9*(12).

Wilkerson, J., & Lang, W. S. (2006). Measuring teaching ability with the Rasch model by scaling a series of product and performance tasks. *Journal of Applied Measurement*, *7*(3), 239–259.

Wilkerson, J. R., & Lang, W. S. (2007). *Assessing teacher dispositions: Five standards-based steps to valid measurement using the DAATS model*. Thousand Oaks, CA: Corwin Press.

Wilkerson, J. R., & Lang, W. S. (2008). Measuring Teacher Dispositions: An application of the Rasch Model to a complex accreditation requirement. *Resources in Education*. (ERIC Document Reproduction Service No. ED502872).

Wilkerson, J. R., & Lang, W. S. (2011). Standards-based teacher dispositions as a necessary and measurable construct. *The International Journal of Educational and Psychological Assessment*, *7*(2), 34–54. Retrieved March 31, 2011 from http://tijepa.books.officelive.com/Documents/A3_V7_2_TIJEPA.pdf

Wilkerson, J. R., Moore, L. L., Lang, W. S., & Zhang, J. (2020). Comparison of students in teacher education from China and the USA: An assessment of dispositions. *International Journal of Learning, Teaching and Educational Research*, *19*(11), 109–126. doi:10.26803/ijlter.19.11.7

Wilson, M. (2023). *Constructing measures: An item response modeling approach (2nd ed.)*. NY: Routledge.

Wolf, R. M. (1988). Rating scales. In J. P. Keeves (Ed.). *Educational research, methodology, and measurement: An international handbook* (pp. 478–481). Oxford, England: Pergamon.

Wolfe, E. W. (2004). Equating and item banking with the Rasch model. In E. Smith & R. Smith (Eds.). *Introduction to Rasch measurement* (pp. 360–390). Maple Grove, MN: JAM Press.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*(4), 33–45, 52. [http://www.rasch.org/memo62.htm]. https://doi.org/10.1111/j.1745-3992.1997.tb00606.x

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. MESA Press. https://www.rasch.org/BTD_RSA/pdf%20[reduced%20size]/Rating%20Scale%20Analysis.pdf

Wright, B. D., & Masters, G. N. (2002). Number of person or item strata: (4*Separation + 1)/3. *Rasch Measurement Transactions*, *16*(3), 888. www.rasch.org/rmt/rmt163f.htm

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. MESA Press. https://www.rasch.org/BTD_RSA/pdf/Best%20Test%20Design.pdf

Charalambos Kollias

9 Placing multiple panel cut scores on the same measurement scale

Abstract: Standard setting entails setting cut scores on an instrument to classify test takers into at least two different categories (i.e., mastery or nonmastery). When cut scores have a life-changing impact on test takers, repeating the standard setting process with a different judge panel and/or different standard setting method is warranted. Not surprisingly, different judge panels and/or different standard setting methods yield different results. In this chapter, we explore how probabilistic models for measurement requiring separable parameters and sufficient estimators allow us to place different standard setting judge panels and/or judges employing different standard setting methods on the same latent scale. The importance of equating test-taker data with judge data is highlighted through examples of equated and unequated analyses. This chapter begins with a brief introduction to standard setting and the Angoff standard setting method, in two of its variants. Then, after an overview of the measurement model used, the instrument, sample, and judge data are described. Finally, the impact of equating or not equating test-taker data with judge data on test-taker recalibrated measurements and pass rates is exemplified. Results show that separate judge panels and different standard setting methods can be estimated on the same measurement scale, enabling a single recommended cut score measurement to be used repeatedly, reducing the need to set cut scores on every instrument. These results demonstrate the persistently repeatable estimation of a unit quantity of a kind that could form the basis for an extension of the International System of Units.

Keywords: standard setting, cut scores, multiple panels, equating, Rasch, Many-faceted Rasch Measurementq (MFRM)

9.1 Introduction

Standard setting is a decision-making process of setting a cut score, a specific point on a scale separating test takers into distinct categories such as pass or fail (Cizek et al., 2004; Hambleton & Eignor, 1978; Kaftandjieva, 2010). The process usually entails recruiting a group of subject matter experts and guiding them to complete a variety of tasks designed to help them recommend a cut score. Cut scores impact test takers, stakeholders, and society as they can hinder test takers from (1) entering university,

Charalambos Kollias, National Foundation for Educational Research, Research Department, Centre for Statistics, Slough, UK

https://doi.org/10.1515/9783111036496-009

(2) getting a promotion, or (3) being licensed to practice a profession (Bergstrom & Lunz, 1999; Grosse & Wright, 1986; Kelley & Schumacher, 1984; Pitoniak & Cizek, 2016; Sireci et al., 2012; Smith et al., 1994). Conducting a standard setting workshop has associated academic and practical challenges ranging from selecting the most appropriate standard setting method to recruiting judges and dealing with practical aspects such as providing judges with food and accommodation (Kollias, 2023). It is such practical challenges that place a financial burden on the awarding organization that may result in cut score workshops not being conducted (Tannenbaum, 2013) and/or replicated at regular intervals (Dunlea & Figueras, 2012).

Cut scores set through different panels and/or standard setting methods yield different standards (Downing et al., 2006; Jaeger, 1989; Kaftandjieva, 2010), limiting measurement comparability, reproducibility, and metrological traceability. Measurement reproducibility refers to the ability to replicate a measurement process and its results on the same or similar objects under specified conditions, which need not necessarily be bound by the same exact external conditions (Mari et al., 2015). Applying this definition in a standard setting context, a test instrument would be the "object" while the standard setting panel would be the "external conditions." At the same time, metrological traceability refers to different instruments measuring the same construct in the same measurement framework (Salzberger, 2023). In a standard setting context, "different instruments" would be the different standard setting methods. Measurement reproducibility and metrological traceability would be achieved when the same recommended raw cut scores yield the same calibrated measurements across standard setting methods and panels. Should this result be obtained, we will have provisionally established a start at developing the kind of an extension of the International System of Units (the SI) Jeckelmann and Edelmaier (2023, p. 2) suggest will be needed "if the SI is to truly live up to its claim to be the universal language for all sciences."

Applying advanced measurement science in the analysis of standard setting judgments has sometimes been viewed as panacea producing measurement reproducibility and metrological traceability when both test-taker scores and judge scores are calibrated together. But separate data analyses typically have their "own origin (zero point)" (Linacre, 2023a, p. 166), because default software settings center each logit scale at the mean of the item difficulties (Andrich & Surla, 2023). Thus, close attention must be invested in substantiating the explicit requirement that test-taker scores and judge scores both are expressed in measurements positioned on the same latent scale (Andrich, 1989; Humpry et al., 2014; Wyse, 2017; Wyse, 2018). This requirement can be satisfied by equating test-taker calibrations with judge data calibrations by anchoring the items to their respective difficulty (Kollias 2023; Wyse, 2017). Such equating will assure metrological traceability of calibrated measurement results (Mari et al., 2023; Pendrill, 2019), especially across standard setting methods. What follows is an illustration of the impact that nonequated and equated calibrated analyses have on test takers' calibrated measurements and pass rates across panels and standard setting methods.

9.2 Methodology

This section provides a brief background description of a commonly used standard setting method and two of its variants to set cut scores on multiple-choice instruments. The test instrument, the test-taker dataset, and the judge datasets are discussed in this section along with the measurement model used for analyzing the judge datasets.

9.2.1 The measurement model

Many-faceted measurement models (Fischer, 1973; Linacre, 1989/1994) expand on probabilistic models for measurements emerging from multiple independent sources over the course of the twentieth century. The family of measurement models (Masters & Wright, 1984; Wright & Mok, 2000) routinely falling under the heading of "Rasch measurement theory" (Rasch, 1960/1980) are conceptually identical with (a) models for paired comparisons previously developed by Zermelo (Linacre, 2000a) and by Bradley-Terry (1952) and Luce (1959) (Andrich, 1988, p. 43; Linacre, 1995), and (b) additive conjoint (Luce & Tukey, 1964) formulations of fundamental measurement (Newby et al., 2009; Wright, 1997). Perhaps most remarkably, C. S. Peirce notably put all the pieces together for a log-odds measurement model in 1878 (Linacre, 2000b).

The significance of these models follows from the point made by Narens and Luce (1986, p. 177) that:

> the development of conjoint structures . . . not only provided a deep measurement analysis of the numerous nonextensive, "derived" structures of physics, but also provided a measurement approach that appears to have applications in the nonphysical sciences and has laid to rest the claim that the only possible basis for measurement is extensive structures.

It is not surprising, then, to find that this class of models is recognized in engineering and the natural sciences for the value obtained (Mari & Wilson, 2014; Mari et al., 2023; Pendrill, 2019; Pendrill & Fisher, 2015). The fundamental implication is that the identification of invariant quantities ought to support development of a new class of SI units.

The basic form of probabilistic measurement requiring separable parameters, sufficient estimators, and an identified structure capable of provisionally achieving what Rasch (1966) referred to as "specific objectivity" (Fischer, 1981; San Martin & Rolin, 2013) incorporates two facets placing test takers and dichotomously scored item difficulties on the same scale. Additional parameters are added as the measurement design expands to include more facets, such as one or more rating scales, tasks, judges, or kinds of items or situations (Bachman, 2004) affecting the interaction between test takers and item difficulty. In order to keep the analysis in the same recalibrated score range (200–800), the recalibration conversion formula used for test-taker score reporting was applied (see Section 9.2.2).

The many-faceted measurement model analyses were conducted using the *FACETS* computer software program (Linacre, 2023b, version 3.86.0). A three-facet model was parameterized: (1) judges; (2) panel; and (3) items. Out of the three facets, two facets were active (judges and items) and one facet (panel) was inactive. In this way, the estimates were based on an interaction between judges and items only, implying that mean cut score measurements would not be influenced by any interactions between active and inactive facets. The three-facet model used in the *FACETS* specification file was the dichotomous model (i.e., Model = ?,?,1-45,D).

The many-faceted measurement model used for analyzing the Yes/No Angoff data can be written as follows:

$$\log\left(\frac{P_{nijk1}}{P_{nijk0}}\right) \equiv B_n - D_i - P_m - D_y \tag{9.1}$$

where P_{nijk1} is the probability of a “Yes” being awarded on item i by judge n, P_{nijk0} is the probability of a “No” being awarded on item i by judge n, B_n is the leniency of judge n, D_i is the difficulty of item i, P_m is the severity of panel m, and D_y is the difficulty of rating a “*Yes*” relative to “*No*.”

Thus, from eq. (9.1), the probability of judge n assigning a “Yes” on item i in panel m rather than a “No” equals the leniency of judge n, minus the difficulty of the item i, minus the severity of panel m, minus the difficulty of assigning a “*Yes*” relative to “*No*” D_y.

To demonstrate that unequated different standard setting methods yield different recalibrated measurements despite using the same raw scores, a further two judge panels (1b and 1c) were created using panel 1 judgments. Panel 1b and panel 1c judgments were created to reflect modified Angoff and extended Angoff judgments, respectively. To retrieve the modified Angoff judgments, panel 1’s Rasch expected scores retrieved from the *FACETS* response file were used as judgments. For example, J01 in the unanchored Yes/No Angoff workshop stated that the minimally expected test taker would answer the first item correctly. In accordance with the Rasch model, J01 had an expected score of 0.9858 for item 1. The expected scores were entered into the modified Angoff dataset twice, once for test-taker success (i.e., R0.9858,J01b,P3,46,1) and once for test-taker failure on the item (i.e., R0.0142,J01b,P3,46,0). For the extended Angoff judgments, the overall individual raw mean cut scores for each judge were used (i.e., J01c,P4,91,30).

The three-facet model used in the specifications for the modified Angoff judgments was a rating scale model (i.e., Model = ?,?,46-90, R45), while the model for the single extended Angoff judgments was a binomial model (i.e., Model = ?,?,91,B45). It should be noted that the modified Angoff judgments can also be “modeled as outcomes of binomial trials [where] the number of independent trials (m) is fixed at ‘100’” (Eckes, 2015, p. 160).

The many-faceted measurement model used for the analyses of the modified Angoff and the extended Angoff data can be written as follows:

$$\log\left(\frac{P_{nijk1}}{P_{nijk-1}}\right) \equiv B_n - D_i - P_m - T_{ik} \tag{9.2}$$

where P_{nijk1} is the probability of k being awarded on item i by judge n, P_{nijk-1} is the probability of $k-1$ being awarded on item i by judge n, B_n is the leniency of judge n, D_i is the difficulty of item i, P_m is the severity of group m, and T_{ik} is the difficulty of assigning k relative to $k-1$.

Thus, from eq. (9.2), the log ratio of the probability of judge n assigning a k on item i and the probability of judge n assigning k–1 on item i equals the leniency of judge n, minus the difficulty of the item i, minus the severity of panel m, minus the difficulty of assigning a k relative to $k-1$ (T_{ik}).

9.2.2 The standard setting method

Cited in the literature as the most widely used test-centered standard setting method, the Angoff method and its variants are the most thoroughly researched methods (Cohen et al., 1999; Plake & Cizek, 2012). Three variants of the Angoff method are commonly used to set cut scores. The first Angoff variant, referred to as the modified Angoff method, originally proposed in a footnote (Angoff, 1971), is used for dichotomously scored items such as multiple-choice items. In this variant, judges are asked to think of 100 minimally competent test takers and decide what proportion of them would answer each item correctly. The second variant, the Yes/No method (Impara & Plake, 1997), also used for dichotomously scored items, requires judges to decide whether a minimally competent test taker would answer each item correctly. In contrast, the third variant, the extended Angoff method (Cizek & Bunch, 2007) requires judges to estimate the average score a minimally competent test taker would get on a polytomously scored item or on a performance-based task. In all three variants, cut scores are usually established by summing up judges raw scores and/or probabilities of success on each item and averaging them. Judgments from all three Angoff methods are calibrated in a common frame of reference and placed on the same measurement scale.

9.2.3 The instrument, judge data, and test-taker dataset

The instrument and the judge data used in the analyses in the next section come from a series of synchronous virtual standard setting workshops (Kollias, 2023). In these workshops, 45 judges were separated into 4 judge panels to set cut scores on 2 equated English language proficiency instruments using the Yes/No Angoff method. Each instrument contained 45 multiple-choice items [15 discrete grammar items, 15 discrete vocabulary items, and 15 reading comprehension items (3 passages × 5 items each)]. As the original instrument had approximately 600 test-taker responses, a dataset of 5,000 test takers was simulated through the WINSTEPS® software program (Linacre, 2023c, version 5.6.0.0). In this way, the impact that different cut score measurements have on test-taker pass rates would not be affected by the instrument's sample size. As the logit (the log-odds unit of measurement) calibrations of the 45 items ranged from −5.2374 (score 0) to 5.2595 (score 45), the measurements were recalibrated into a positive linear scale (Linacre, 2023a), ranging from 200 to 800. The simulated logit measurements were linearly transformed into rescaled positive measurements by multiplying each logit by a predefined user-scaled unit (using the WINSTEPS "uscale" command) and adding the items' mean difficulty [Rescaled measurement = (logit measurement × uscale) + rescaled item mean difficulty]. For example, a logit of −0.0552 (score of 22) was transformed into a rescaled measurement of 496 (−0.0552 × 57.1597) + 499.3684 = 496). Table 9.1 displays both the logits and the rescaled measurements for each of the 45 score points.

Table 9.1: Table of measurements on test of 45 items.

Score	Logit (S.E.)	Rescaled measurement[1] (S.E.)	Score	Logit (S.E.)	Rescaled measurement[1] (S.E.)
0	−5.2 (1.8)	200 (105)	23	0.0 (0.3)	502 (18)
1	−4.0 (1.0)	270 (58)	24	0.1 (0.3)	508 (18)
2	−3.3 (0.7)	312 (42)	25	0.2 (0.3)	513 (18)
3	−2.8 (0.6)	337 (35)	26	0.3 (0.3)	519 (18)
4	−2.5 (0.5)	355 (30)	27	0.5 (0.3)	525 (18)
5	−2.3 (0.5)	370 (28)	28	0.6 (0.3)	531 (19)
6	−2.1 (0.4)	382 (26)	29	0.7 (0.3)	537 (19)
7	−1.9 (0.4)	393 (24)	30	0.8 (0.3)	544 (19)
8	−1.7 (0.4)	403 (23)	31	0.9 (0.3)	550 (19)
9	−1.5 (0.4)	412 (22)	32	1.0 (0.3)	557 (20)
10	−1.4 (0.4)	420 (21)	33	1.1 (0.4)	564 (20)
11	−1.3 (0.4)	428 (21)	34	1.3 (0.4)	571 (21)
12	−1.1 (0.4)	435 (20)	35	1.4 (0.4)	579 (21)
13	−1.0 (0.3)	442 (20)	36	1.5 (0.4)	587 (22)
14	−0.9 (0.3)	448 (19)	37	1.7 (0.4)	596 (23)
15	−0.8 (0.3)	455 (19)	38	1.9 (0.4)	606 (24)
16	−0.7 (0.3)	461 (19)	39	2.1 (0.5)	617 (26)
17	−0.6 (0.3)	467 (19)	40	2.3 (0.5)	629 (28)
18	−0.5 (0.3)	473 (18)	41	2.5 (0.5)	644 (31)
19	−0.4 (0.3)	479 (18)	42	2.9 (0.6)	663 (35)
20	−0.3 (0.3)	485 (18)	43	3.3 (0.7)	688 (42)
21	−0.2 (0.3)	490 (18)	44	4.0 (1.0)	730 (58)
22	−0.1 (0.3)	496 (18)	45	5.3 (1.8)	800 (105)

[1] Rescaled measurement = (57.1597 × logit) + 499.3684.

Two judge panels (panel 1 and panel 2) were created using the judge data from the virtual standard setting workshops. Each panel consisted of data from 10 judges, the minimum number of judges recommended when using an Angoff standard setting method (Brandon, 2004). The Yes/No Angoff judge data for each panel were selected so that the difference in the raw mean cut scores between the two panels would be exactly one raw score point. Both panel data were analyzed using a many-faceted measurement model, a measurement model that has become more prevalent in analyzing standard setting data in the last decade (e.g., see Kollias, 2023; Peabody & Wind, 2019; Roberts et al., 2017; Wu & Tan, 2016).

9.3 Cut score analysis

This section compares and contrasts the unequated with the equated many-faceted measurement analyses between different judge panels and across different standard setting methods. The rescaled cut score measurements derived from the equated and unequated analyses are evaluated in terms of their impact on test-taker pass rates. In this section, a new way of analyzing one-item extended Angoff judge ratings through many-faceted measurement is also presented.

9.3.1 Unequated analysis between panels

The first set of analyses entailed analyzing panel 1 and panel 2 Yes/No Angoff datasets separately. Table 9.2 reports the separate many-faceted measurement analyses. Column 1 (judge) displays the judge code, while column 2 (score) and column 3 (rescaled measurement) report each judge's raw mean cut score and rescaled cut score measurement with its corresponding standard errors in parenthesis, respectively. Panel 1 had a raw mean cut score of 27.40 which was transformed into a rescaled cut score measurement of 538, while panel 2 had a raw mean cut score of 28.40, a higher raw mean cut score, but had a lower rescaled mean cut score measurement of 533. If both analyses were on the same latent scale, we would expect a higher raw mean cut score to yield a higher rescaled mean cut score measurement.

Further examination of individual recalibrated mean cut score measurements revealed in greater detail the discrepancy observed between the panels' raw mean cut scores and their rescaled measurements. For example, in panel 1, judges J02, J08, and J09 had a raw mean cut score of 27 which was transformed into a recalibrated mean cut score measurement of 532. In contrast, in panel 2, judges J11 and J12 who also had a raw mean cut score of 27 had a recalibrated mean cut score measurement of 518. A raw mean cut score of 30 yielded a recalibrated mean cut score measurement of 573 for panel 1 (J01) and 547 for panel 2 (J15), and a raw mean cut score of 31 transformed

Table 9.2: Panel 1 Yes/No Angoff rescaled mean cut score measurements (unequated).

Panel 1			Panel 2		
Judge	**Score**	**Rescaled measurement (S.E.)**	**Judge**	**Score**	**Rescaled measurement (S.E.)**
J01	30	573 (29)	J11	27	518 (23)
J02	27	532 (28)	J12	27	518 (23)
J03	30	573 (29)	J13	34	591 (26)
J04	26	518 (28)	J14	22	473 (23)
J05	22	465 (28)	J15	30	547 (24)
J06	30	573 (29)	J16	26	509 (23)
J07	24	492 (28)	J17	31	557 (24)
J08	27	532 (28)	J18	31	557 (24)
J09	27	532 (28)	J19	30	547 (24)
J10	31	588 (28)	J20	26	509 (24)
Mean	27.40	538 (28)	Mean	28.40	533 (24)

into 588 for panel 1 (J10) and 557 for panel 2 (J17 and J18). The comparison evinced that panel 2 rescaled measurements were consistently lower than those of panel 1.

When comparing the same individual rescaled mean cut score measurements (i.e., 27, 30, and 31) with the rescaled test-taker measurements (see Table 9.1), it is apparent that not only are the two judge panel analyses not equated with one another, but neither analysis is equated with test-taker ability. Figure 9.1 illustrates the difference in unequated individual recalibrated mean cut score measurements between panels and test-taker ability.

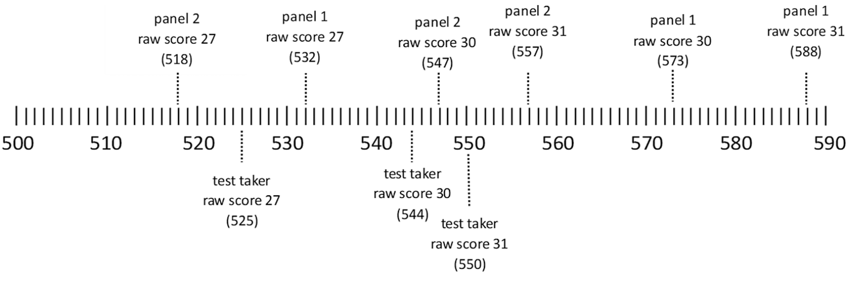


Figure 9.1: Comparison of unequated recalibrated cut score measurements between panels.

9.3.2 Unequated analyses across panels

The separate many-faceted measurement analyses revealed that the rescaled mean cut score measurements for each standard setting method were not the same despite the same raw scores being used in the analysis. Table 9.3 reports the recalibrated mean cut score measurements for all three methods, and the rescaled mean cut score measurements ranged from 525 to 538. The same findings were observed when individual recalibrated mean cut score measurements were compared across methods. For example, J05, who recommended the lowest cut score of all 10 judges in panel 1, had a raw mean cut score of 22, which was transformed into a rescaled measurement of 465 (panel 1), 480 (panel 1b), and 497 (panel 1c) for each standard setting method, respectively. However, test takers who scored 22 out of 45 received a recalibrated measurement of 496 (see Table 9.1). Figure 9.2 illustrates the difference in unequated panel recalibrated mean cut score measurements across standard setting methods and test-taker ability.

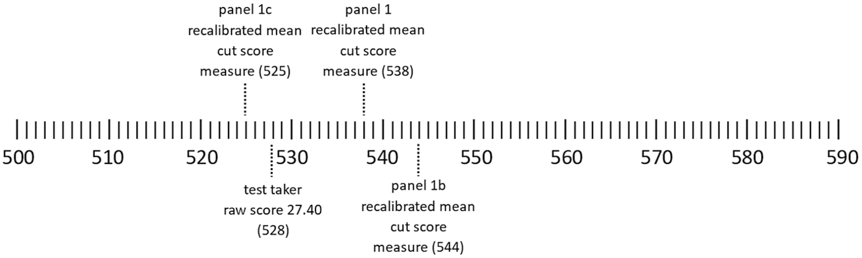

panel 1c recalibrated mean cut score measure (525)
panel 1 recalibrated mean cut score measure (538)
500
510
520
530
540
550
560
570
580
590
test taker raw score 27.40 (528)
panel 1b recalibrated mean cut score measure (544)

Figure 9.2: Comparison of unequated recalibrated cut score measurements across standard setting methods.

Table 9.3: Panel 1 rescaled mean cut score measurements by standard setting method (unequated).

Judge	Score	Yes/No Angoff (panel 1)	Modified Angoff (panel 1b)	Extended Angoff (panel 1c)
		Rescaled measurement (S.E.)	Rescaled measurement (S.E.)	Rescaled measurement (S.E.)
J01	30	573 (29)	576 (27)	539 (18)
J02	27	532 (28)	539 (29)	523 (17)
J03	30	573 (29)	576 (27)	539 (18)
J04	26	518 (28)	527 (26)	517 (17)
J05	22	465 (28)	480 (25)	497 (17)
J06	30	573 (29)	576 (27)	539 (18)
J07	24	492 (28)	503 (26)	507 (17)
J08	27	532 (28)	539 (26)	523 (17)
J09	27	532 (28)	539 (26)	523 (17)
J10	31	588 (28)	589 (27)	545 (18)
Mean	27.40	538 (28)	544 (26)	525 (18)

The impact that unequated analyses have on test-taker pass rates is evident in Table 9.4. Panel 1 had a raw mean cut score of 27.40 yielding a pass rate ranging from 47.64% (panel 1b) to 63.18% (panel 1c). For panel 2, a raw mean cut score of 28.40 yielded a higher pass rate of 56.04%, greater than those of panel 1 and panel 1b.

Table 9.4: Panel pass rates (unequated).

Panel	Standard setting method	Count	Mean raw cut score	Recalibrated measurement (S.E.)	Pass rate
1	Yes/No Angoff	45	27.40	538 (28)	51.78%
2	Yes/No Angoff	45	28.40	533 (24)	56.04%
1b	Modified Angoff	45	27.40	544 (26)	47.64%
1c	Extended Angoff	1	27.40	525 (18)	63.18%

9.3.3 Equating test-taker data and judge data

The next set of analyses entailed equating the panel data and standard setting methods by anchoring the items to their respective difficulty calibrations. Table 9.5 reports the item anchored rescaled cut score mean calibrations. The analyses revealed that panel 1 raw mean cut score of 27.40 was transformed into a rescaled measurement of 528, while panel 2 raw mean cut score of 28.40 was transformed into a rescaled measurement of 534. By anchoring items to their respective difficulty measurements, both panel judgments and test-taker data are placed on the same latent scale. Consequently, it can be observed that a raw score of 30 transformed into a rescaled measurement of 544 for both panels, thus matching the rescaled score point measurements of the instrument (see Table 9.1). Figure 9.3 illustrates the rescaled mean cut score measurements across equated panels and test-taker ability.

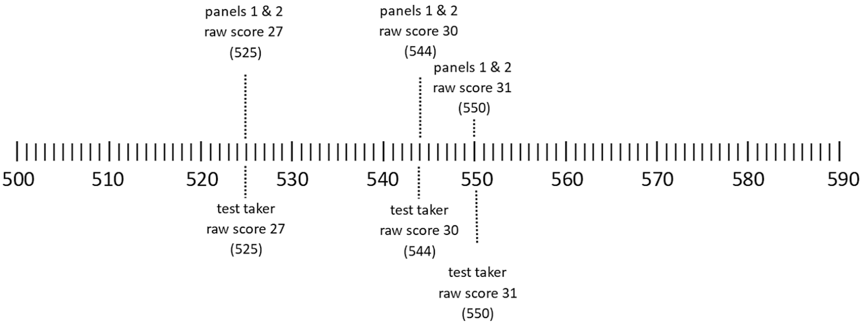


Figure 9.3: Comparison of equated recalibrated cut score measurements between panels.

9.3.4 Equated analyses across standard setting methods

Comparing anchored analysis across standard setting methods revealed that the extended Angoff rescaled measurements (panel 1c) were not on the same latent scale as the other two Angoff variants (see Table 9.6). Panel 1 and panel 1b rescaled measure-

Table 9.5: Panel 1 Yes/No Angoff rescaled mean cut score measurements (equated).

Panel 1			Panel 2		
Judge	**Score**	**Rescaled measurement (S.E.)**	**Judge**	**Score**	**Rescaled measurement (S.E.)**
J01	30	544 (19)	J11	27	525 (18)
J02	27	525 (18)	J12	27	525 (18)
J03	30	544 (19)	J13	34	571 (21)
J04	26	519 (18)	J14	22	496 (18)
J05	22	496 (18)	J15	30	544 (19)
J06	30	544 (19)	J16	26	519 (18)
J07	24	508 (18)	J17	31	550 (19)
J08	27	525 (18)	J18	31	550 (19)
J09	27	525 (18)	J19	30	544 (19)
J10	31	550 (19)	J20	26	519 (18)
Mean	27.40	528 (19)	Mean	28.40	534 (19)

ments matched each other as well as the rescaled score point measurements of the instrument (see Table 9.1). However, the anchoring of items to their respective difficulty calibrations yielded the same rescaled mean cut score measurement for the extended Angoff method (panel 1c) as those in panel 1c unequated analysis while only minor differences in individual judge recalibrated mean cut score measurements were observed.

Consequently, a new dataset (panel 1d) was created by transforming the extended Angoff raw cut scores into percentages and entering the data as they were entered in the modified Angoff dataset. For example, J01 had a raw score of 30 out of 45, which is equivalent to 0.6667. Consequently, for J01, the 0.6667 proportion of test-taker item success (i.e., R0.6667, J01, P5, 92, 1; R0.6667, J01, P5, 93, 1; . . .; R0.6667, J01, P5, 136, 1) and its corresponding proportion 0.3333 for test-taker failure (i.e., R0.3333, J01, P5, 92, 0; R0.3333, J01, P5, 93, 0; . . .; R0.3333, J01, P5, 136, 0) were entered in the dataset for all 45 items in panel 1d dataset. The facets model used for panel 1d analysis was the rating scale model (i.e., Model = ?, ?, 92-136, R45).

The dataset and the new analysis model were confirmed with Dr. John Michael Linacre (July 2023, personal communication), who suggested an alternative way of entering the data for analysis. The dataset could be created by combining how the Yes/No Angoff data and the modified Angoff data were entered for analysis. For example, for J01, the data would be entered as follows:

R0.6667,J01d,P5,92-136,1,1,1, . . ., 1,1,1; (i.e., “1” entered 45 times for test-taker success) and
R0.3333,J01d,P5,92-136,0,0,0, . . ., 0,0,0; (i.e., “0” entered 45 times for test-taker failure)

Panel 1d analysis yielded rescaled mean cut score measurements that matched those of the other two methods (Yes/No Angoff and the modified Angoff). Figure 9.4 illus-

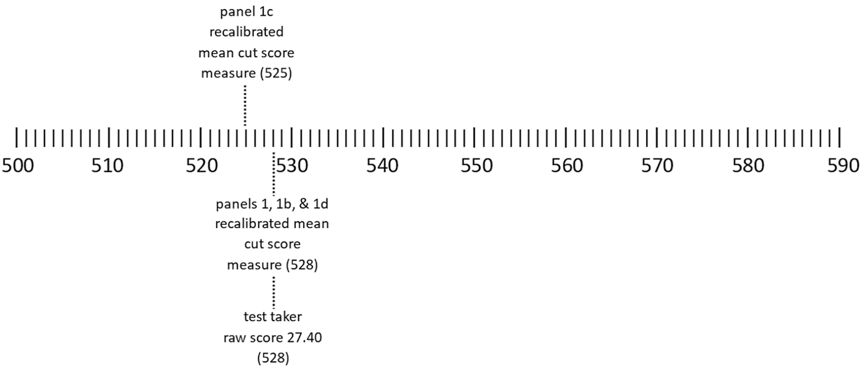


Figure 9.4: Comparison of equated recalibrated cut score measurements across standard setting methods.

Table 9.6: Panel A rescaled mean cut score measurements by standard setting method (equated).

Standard setting method		**Yes/No Angoff (panel 1)**	**Modified Angoff (panel 1b)**	**Extended Angoff (panel 1c)**	**Extended Angoff (panel 1d)**
Judge	**Score**	**Rescaled measurement (S.E.)**	**Rescaled measurement (S.E.)**	**Rescaled measurement (S.E.)**	**Rescaled measurement (S.E.)**
J01	30	544 (19)	544 (19)	539 (18)	544 (19)
J02	27	525 (18)	525 (18)	522 (17)	525 (18)
J03	30	544 (19)	544 (19)	539 (18)	544 (19)
J04	26	519 (18)	519 (18)	517 (17)	519 (18)
J05	22	496 (18)	496 (18)	496 (17)	496 (18)
J06	30	544 (19)	544 (19)	539 (18)	544 (19)
J07	24	508 (18)	508 (18)	507 (17)	508 (18)
J08	27	525 (18)	525 (18)	522 (17)	525 (18)
J09	27	525 (18)	525 (18)	522 (17)	525 (18)
J10	31	550 (19)	550 (19)	544 (18)	550 (19)
Mean	27.40	528 (19)	528 (19)	525 (18)	528 (19)

trates the panel recalibrated mean cut score measurements across equated standard setting methods and test-taker ability.

The final analysis entailed calculating the pass rates for each equated panel. Table 9.7 displays the corresponding pass rates for each panel. As the items were anchored to their respective difficulties, the pass rates for panels 1 and 2 were as expected. Compare the panel 2 pass rate with that of panel 1, and panel 2's lower pass rate followed from its higher rescaled mean cut score measurement. Panels 1, 1b, and 1d had the same pass rate, illustrating that equating the three different standard setting methods was successful. Apart from panel 1c measurements (analyzed through the binomial method), all the other panel recalibrated measurements were now on the same measurement scale.

Table 9.7: Panel pass rates (equated).

Group	Standard setting method	Count	Mean cut score	Rescaled cut score measurement	Pass rate
1	Yes/No Angoff	45	27.4	528 (19)	59.78%
2	Yes/No Angoff	45	28.4	533 (24)	56.04%
1b	Modified Angoff	45	27.4	528 (19)	59.78%
1c	Extended Angoff	1	27.4	525 (18)	63.18%
1d	Extended Angoff	1	27.4	528 (19)	59.78%

9.4 Discussion

In this chapter, we discussed a method of equating test-taker data with judge data to assure measurement reproducibility and metrological traceability of calibrated measurements. By equating test-taker data with judge data by anchoring item difficulty, we were able to place both separate panels and different standard setting methods on the same measurement scale. This, in turn, will allow awarding organizations to use the recommended cut score measurements on subsequent equated versions of their test instruments thus reducing the need to set cut scores on every instrument.

We also demonstrated that unequated analyses yielded different results, sometimes contradicting Rasch's (1966) specific objectivity principle, denoting that higher raw cut scores yield higher recalibrated measurements instead of the expected lower measurements. The impact that unequated analyses have on test-taker pass rates may have detrimental effects on test takers, especially in cases of licensure. The equated analyses demonstrated specific objectivity, as the same raw score on the same set of items produced the same measurements.

Thus, it can be concluded that only when cut scores have measurement reproducibility and metrological traceability can test score interpretations and test-taker classification based on calibrated measurements be valid.

Validity of this kind is established via repeated demonstrations that (a) experimental tests across samples and instruments reproduce the same unit quantity, and that (b) theoretical explanations of variation successfully predict the respective measurements and calibrations to fit-for-purpose tolerances (Pendrill, 2019). Plainly further research will be required if we are to extend the SI into new domains at a level satisfying Feynman's (1988) criterion for grasping the situation: "What I cannot create, I do not understand."

References

Andrich, D. (1988). *Sage University Paper Series on Quantitative Applications in the Social Sciences. Vol. series no. 07-068: Rasch models for measurement.* Sage Publications.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J. A. Keats, R. Taft, R. A. Heath, & S. H. Lovibond (Eds.). *Mathematical and theoretical systems: Proceedings of the 24th International Congress of Psychology of the International Union of Psychological Science* (Vol. 4, pp. 7–16). Elsevier Science Publishers.

Andrich, D., & Surla, D. (2023). Equating measuring instruments in the social sciences: Applying measurement principles of the natural sciences. In W. P. Fisher Jr. & S. J. Cano (Eds.). *Person-centred outcome metrology: Principles and applications for high-stakes decision making* (pp. 195–226). Springer Nature Switzerland AG. doi:10.1007/978-3-031-07465-3_8

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.). *Educational measurement* (pp. 508–600). 2nd ed. Washington: American Council of Education.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olson-Buchanan (Eds.). *Innovations in computerized assessment* (pp. 67–91). Lawrence Erlbaum Associates, Inc., Publishers.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of pair comparisons. *Biometrika*, *63*, 324–345.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, *17*(1), 59–88. doi:10.1207/s15324818ame1701_4

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage Publications.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, *23*(4), 31–50. doi:10.1111/j.1745-3992.2004.tb00166.x

Downing, S. M., Tekian, A., & Yudkowsky, R. (2006). Research methodology: Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teaching and Learning in Medicine*, *18*(1), 50–57. doi:10.1207/s15328015tlm1801_11

Dunlea, J., & Figueras, N. (2012). Replicating results from a CEFR test comparison project across continents. In D. Tsagari & C. Ildikó (Eds.). *Collaboration in language testing and assessment* (Vol. 26, pp. 31–45). Peter Lang.

Eckes, T. (2015). *Introduction to Many-facet Rasch measurement: Analysing and evaluating rater-mediated assessments* (2nd Revised and updated edition). Frankfurt: Peter Lang.

Feynman, R. (1988, February 15). Richard Feynman's blackboard at the time of his death. In CalTech Image Archive. Retrieved 31 December 2019, from California Institute of Technology: http://archives-dc.library.caltech.edu/islandora/object/ct1%3A551

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. Psychometrika, 46(1), 59–77.

Grosse, M. E., & Wright, B. D. (1986). Setting, evaluating, and maintaining certification standards with the Rasch model. *Evaluation & the Health Professions*, *9*(3), 267–285.

Hambleton, R., & Eignor, D. R. (1978). Competency test development, validation, and standard-setting. *Paper presented at the Minimum Competency Testing Conference of the American Education Research Association* (pp. 1–50). Washington, DC. Retrieved from http://files.eric.ed.gov/fulltext/ED206725.pdf

Humphry, S., Heldsinger, S., & Andrich, D. (2014). Requiring a consistent unit of scale between the responses of students and judges in standard setting. *Applied Measurement in Education*, *27*, 1–18. doi:10.1080/08957347.2014.859492

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, *34*(4), 353–366. doi:10.1111/j.1745-3984.1997.tb00523.x

Jaeger, R. M. (1989). The certification of student competence. In R. L. Linn (Ed.). *Educational measurement*. 3rd ed. Macmillan.

Jeckelmann, B., & Edelmaier, R. (Eds.). (2023). *Metrological infrastructure*. (De Gruyter Series in Measurement Sciences, K.-D. Sommer & T. Fröhlich, Eds.). De Gruyter Oldenbourg. https://doi.org/10.1515/9783110715835

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading in EFL*. Arnhem: Cito. Retrieved from http://www.ealta.eu.org/documents/resources/FK_second_doctorate.pdf

Kelley, P. R., & Schumacher, C. F. (1984). The Rasch model: Its use by the National Board of Medical Examiners. *Evaluation & the Health Professions*, *7*(4), 443–454.

Kollias, C. (2023). *Virtual standard setting: Setting cut scores*. Peter Lang. doi:10.3726/b20407

Linacre, J. M. (1989/1994). *Many-Facet Rasch measurement*. Chicago: Mesa Press.

Linacre, J. M. (1995). Paired comparisons with ties: Bradley-Terry and Rasch. *Rasch Measurement Transactions*, *9*(2), 425. http://www.rasch.org/rmt/rmt92d.htm

Linacre, J. M. (2000a). Almost the Zermelo model? *Rasch Measurement Transactions*, *14*(2), 754. http://www.rasch.org/rmt/rmt142k.htm

Linacre, J. M. (2000b). Was the Rasch model almost the Peirce model? *Rasch Measurement Transactions*, *14*(3), 756–757. http://www.rasch.org/rmt/rmt143b.htm

Linacre, J. M. (2023a). Advancing the metrological agenda in the social sciences. In *Person-centred outcome metrology: Principles and applications for high stakes decision making* (pp. 165–193). Springer Nature Switzerland AG. doi:10.1007/978-3-031-07465-3_7

Linacre, J. M. (2023b). Facets (Many-Facet Rasch Measurement) computer program (Version 3.86.0) [Computer software]. Retrieved from www.winsteps.com

Linacre, J. M. (2023c). Winsteps® (Version 5.6.0.0) [Computer Software]. www.winsteps.com

Luce, R. D. (1959). *Individual choice behavior*. Wiley.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new kind of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27.

Mari, L., Carbone, P., & Petri, D. (2015). Fundamentals of hard and soft measurement. In A. Ferrero, D. Petri, P. Carbone, & M. Catelani (Eds.). *Modern measurements: Fundamentals and applications* (pp. 203–262). John Wiley & Sons, Inc.

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement*, *51*, 315–327. http://www.sciencedirect.com/science/article/pii/S0263224114000645

Mari, L., Wilson, M., & Maul, A. (2021). *Measurement across the sciences; Developing a shared concept system for measurement*. 2nd ed. Springer Nature Switzerland AG.

Mari, L., Wilson, M., & Maul, A. (2023). Measurement across the sciences: Developing a shared concept system for measurement, 2nd ed. Springer Series in Measurement Science and Technology). Springer.https://link.springer.com/book/10.1007/978-3-031-22448-5

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika*, *49*(4), 529–544. doi:10.1007/BF02302590

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, *99*(2), 166–180.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*(4), 33–45, 52. https://doi.org/10.1111/j.1745-3992.1997.tb00606.x

Newby, V. A., Conner, G. R., Grant, C. P., & Bunderson, C. V. (2009). The Rasch model and additive conjoint measurement. *Journal of Applied Measurement*, *10*(4), 348–354.

Peabody, M. R., & Wind, S. A. (2019). Exploring the influence of judge proficiency on standard-setting judgments. *Journal of Educational Measurement*, *56*(1), 101–120. doi:10.1111/jedm.12202

Pendrill, L. R. (2019). *Quality assured measurement: Unification across social and physical sciences*. Springer Series in Measurement Science and Technology. Springer. https://link.springer.com/book/10.1007/978-3-030-28695-8

Pendrill, L., & Fisher, W. P., Jr. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement*, *71*, 46–55. http://dx.doi.org/10.1016/j.measurement.2015.04.010

Pitoniak, M. J., & Cizek, G. J. (2016). Standard setting. In C. S. Wells & M. Faulkner-Bond (Eds.). *Educational measurement: From foundations to future* (pp. 38–61). The Guildford Press.

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, the extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.). *Setting performance standards: Foundations, methods, and innovations* (pp. 181–199). 2nd ed. Routledge.

Rasch, G. (1960/1980). *Probabilistic models in some intelligence and attainment test*. University of Chicago Press.

Rasch, G. (1966). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.). *Readings in mathematical social science* (pp. 89–108). Science Research Associates. https://www.rasch.org/memo19662.pdf

Roberts, W. L., Boulet, J., & Sandella, J. (2017). Comparison study of judged clinical skills competence from standard setting ratings generated under different administration conditions. *Advances in Health Sciences*, *22*, 1279–1292. doi:10.1007/s10459-017-9766-1

Salzberger, T. (2023). Addressing traceability in social measurement establishing a common metric for dependence. In W. P. Fisher Jr. & S. Cano (Eds.). *Person-centred outcome metrology: Principles and applications for high stakes decision making* (pp. 227–267). Nature Switzerland AG: Springer. doi:10.1007/978-3-031-07465-3_9

San Martin, E., & Rolin, J. M. (2013). Identification of parametric Rasch-type models. *Journal of Statistical Planning and Inference*, *143*(1), 116–130.

Sireci, S. G., Randall, J., & Zenisky, A. (2012). Setting valid performance standards on educational tests. *CLEAR Exam Review*, *23*(2), 18–27.

Smith, R. M., Julian, E., Lunz, M., Stahl, J., Schulz, M., & Wright, B. D. (1994). Applications of conjoint measurement in admission and professional certification programs. *International Journal of Educational Research*, *21*(6), 653–664.

Tannenbaum, R. J. (2013). Setting standards on the TOEIC(R) listening and reading test and the TOEIC(R) speaking and writing tests: A recommended procedure. In *The research foundation for the TOEIC tests: A compendium of studies* (Vol. II, pp. 8.1–8.12). Educational Testing Service.

Wright, B. D. (1997). A history of social science measurement. Educational Measurement: Issues and Practice, 16(4), 33–45, 52.

Wright, B. D., & Mok, M. (2000). Understanding Rasch measurement: Rasch models overview. *Journal of Applied Measurement*, *1*(1), 83–106.

Wu, S. M., & Tan, S. (2016). Managing rater effects through the use of FACETS analysis: The case of a university placement test. *Higher Education Research & Development*, *35*(2), 380–394. doi:10.1080/07294360.2015.1087381

Wyse, A. E. (2017). Five methods for estimating Angoff cut scores with IRT. *Educational Measurement: Issues and Practice*, *36*(4), 16–27. doi:10.1111/emip.12161

Wyse, A. E. (2018). Equating Angoff standard-setting ratings with the Rasch model. *Measurement: Interdisciplinary Research and Perspectives*, *16*(3), 181–194. doi:10.1080/15366367.2018.1483170

Greg Sampson, Yin Burgess, Nadine McBride, and Brent Stevenor

10 A many-faceted measurement modeling approach for informing test specifications: practical guidance from the National Registry of Emergency Medical Technicians

Abstract: It is common practice for licensure and certification organizations to routinely conduct research studies, referred to as Practice Analyses, to identify the essential knowledge and skills required for competent practice by the licensed or certified community as well as to define the job role that is being measured by the licensure or certification examination. The National Registry of Emergency Medical Technicians follows multiple steps and examines a collection of various data (e.g., observational studies, practitioner interviews, Subject Matter Expert panel meetings) during their stringent examination development process. A practice analysis supports the alignment of the test items to the knowledge and skills required for performance in the role and serves as a major argument supporting the content validity of the examination. As part of the practice analysis study for a new examination, data for the testing specification development were collected and subsequently analyzed. Using a *Task List* approach toward the test content development, the Registry aligns the tasks and specific job behaviors to their overarching descriptors and classifications defined as (*Domains*). To represent every task within a domain on every test for every candidate, a probabilistic modeling measurement approach was implemented, thereby taking into account missing observations. The Rasch-based multifaceted measurement model provides a robust tool to precisely distinguish aspects of high stakes licensure and certification assessments, enabling us to incorporate all needed components within a singular and unified measurement activity offering simplicity, elegance, and parsimony. General procedures of each step are detailed and explained; sample Facets codes are provided to aid practitioners in applying this approach. A summary of model performance and analysis results are listed and discussed.

Keywords: practice analysis, job analysis, multifaceted measurement model, test specifications, content validity, licensure and certification

Note: On behalf of the National Registry of EMTs.

Greg Sampson, **Yin Burgess**, **Nadine McBride**, **Brent Stevenor**, The National Registry of EMTs, MI, USA

https://doi.org/10.1515/9783111036496-010

10.1 Introduction

This chapter illustrates an application of multifaceted measurement modeling to the development of testing specifications used in licensure and certification examinations administered by the National Registry of Emergency Medical Technicians. Two important concepts set the context:

1. The Registry uses a *Task List* approach toward its test content development. *Tasks* are specific job behaviors. Comprehensive knowledge of these job behaviors is imperative to the successful practice of emergency medical services (EMSs). Examples from the test discussed here include tasks such as (a) manage interventions to special populations; (b) protect self, other responders, patient, public, and the emergency scene from existing and potential hazards; and (c) assess the patients' airway.
2. Tasks are aligned to *Domains*. These are overarching descriptors and classifications of tasks. Specific tasks are grouped by these domains. The domains in the test discussed here include scene size-up and safety, primary assessment, secondary assessment, patient treatment and transport, and operations.

Both administratively and psychometrically speaking, representing every task within a domain on every test for every candidate would be an unmanageably cumbersome undertaking. However, this can be accomplished using a probabilistic modeling measurement framework capable of taking missing observations into account. Many-faceted measurement modeling provides powerful tools for test developers and subject matter experts (SMEs) when tests need to be weighted according to their tasks and domains. Methods are provided here.

The example of a practical implementation of measurement modeling we present assumes readers possess a basic understanding of the relevant theory (Andrich & Marais, 2019; Bond et al., 2021; Linacre, 1994; Linacre et al., 1994; Myford & Wolfe, 2009; Smith et al., 1994; Wilson, 2023). The aim is to clearly articulate what can be accomplished in the measurement context of a job analysis, especially as it applies to the weighting of tests. Multifaceted measurement modeling is useful in placing complex job analysis results on a common, interpretable scale. We will take up the following topics:

1. a quick overview of multifaceted measurement modeling concepts;
2. measurement modeling in the context of job analysis;
3. a practical set of instructions for completing a multifaceted modeling project; and
4. a real-life example to show how the approach can be implemented with the relevant software (Linacre, 2023).

10.2 Multifaceted measurement modeling

Multifaceted models for probabilistic conjoint measurement provide robust and easily available tools applicable in a wide variety of assessment situations. These models are part of the family of models typically gathered together under the heading of "the Rasch model" and which are well-established as being identical with the Zermelo and Bradley-Terry models of paired comparisons, and with the Luce choice axiom (Andrich, 1988, p. 43; Linacre, 1995, 2000). They constructively augment the sound measurement principles expounded by Thurstone and Guttman (Andrich, 1978; Engelhard, 2008). The parameter separation and specific objectivity ideals set using these models conform to additive conjoint formulations of fundamental measurement (Newby et al., 2009; Wright, 1997).

Philosophically, these measurement models have been connected to the origins of geometry and the history of mathematical thinking going back to Plato (Fisher, 2003a/b, 2004). It should also be noted that physicists and engineers responsible for the management and improvement of the SI units (the "metric system") have gone on the record saying that the approach represented by these models "is not simply a mathematical or statistical approach, but [is] instead a specifically metrological approach to human-based measurement" (Pendrill, 2014, p. 26; also see Mari & Wilson, 2014, p. 326; Mari et al., 2023; Pendrill, 2019; Pendrill & Fisher, 2015). We then feel that it is time to stop referring to probabilistic measurement modeling concepts and methods in ways that counterproductively and unnecessarily restrict their scope within a narrowly limited domain. We aim here to join with the other chapters in this book to begin articulating a broader sense of measurement modeling accessible to a wider range of researchers working in different fields.

Multifaceted approaches provide exacting and precise methods for distinguishing aspects of high stakes certification and licensure assessments that might otherwise be confounded (Myford & Wolfe, 2009; Smith et al., 1994; Tavakol & Pinner, 2019; Warner et al., 2020). The multifaceted approach incorporates these known components within a singular and unified measurement activity that, like other models of this kind, offers otherwise unavailable degrees of simplicity, elegance, and parsimony (Cano et al., 2016). Or, put another way, multifaceted models can include explanatory variables predicting person and item locations in facets predicting how different raters, differing scales, person-level attributes, testing conditions, test structures, task components, and so on can be studied in relation to one another (Bond et al., 2021; De Boeck & Wilson, 2004; Eckes, 2015; Fischer, 1973; Linacre, 1994; Melin et al., 2021; Stenner et al., 2013; Wind & Hua, 2022).

Multifaceted models are often described as expanding on dichotomous, partial credit, and rating scale models by

1. incorporating repeated measurements (e.g., pre- and posttreatment measurements of academic ability within an experimental design);
2. including evaluations of potentially confounding components (e.g., rater by time effects);

3. adding in feature analyses (e.g., multiple subscales comprising an overall common scale); or
4. estimating variance components with multiple variables (as in G-theory; Li et al., 2021) while simultaneously producing standard measurement quality evaluations such as infit and outfit.

10.3 A use-case for multifaceted measurement modeling

The Registry is the EMS Credentialing organization for the United States. There are over 300,000 EMS certification holders across the nation. As a large-scale assessment and credentialing program, the Registry implements contemporary, research-based, defensible methods of test design and score reporting within four different operational testing programs. These testing programs have been increasingly based on principles of probabilistic conjoint measurement modeling since the 1970s (Bergstrom & Lunz, 1999; Kelley & Schumacker, 1984; Grosse & Wright, 1986; Smith et al., 1994; Tavakol & Pinner, 2019; Warner et al., 2020).

The Registry recently completed a job analysis for the Nationally Registered Emergency Medical Responder Program and the Nationally Registered Emergency Medical Technician Program. The analytical plan included applying a multifaceted measurement model to the collected data, extending the work of Wang and Stahl (2012). One of the primary outcomes of any job analysis related to examination development for certification is the estimation of a metric that captures *Task Importance* (sometimes referred to as skill importance, ability importance, etc.) relative to the content of the assessment. As noted above, these tasks are classified into overarching performance domains that are used to guide content weights on every administered test. With this approach, the importance metric directly informs test content decisions and this metric stems directly from job analysis.

10.3.1 Job analysis

Job analysis is the process of identifying the tasks that are performed within a job and the knowledge, skills, abilities, and other characteristics (KSAOs) that are required to successfully perform the job tasks. A primary outcome of a job analysis is an updated job description that indicates what tasks are performed within a job as well as how and why they are performed. Also, by identifying the KSAOs required to perform the job tasks, a job analysis allows for the design of selection and certification exams that assess relevant KSAOs.

Job analysis is a rigorous process aimed at identifying the work- (i.e., tasks) and worker-oriented (i.e., KSAOs) elements of a job. A traditional job analysis will consist of multiple steps such as gathering information from existing sources (e.g., previous job descriptions), conducting job observations, interviewing SMEs, and administering structured questionnaires to SMEs. The data gathered from existing sources, observations, and interviews is used to create a large list of tasks and KSAOs that are included in the questionnaire. SMEs then complete the structured questionnaire in which they rate the frequency and importance (i.e., *criticality* in the field of emergency medicine) of each task and KSAO. Once the critical tasks and KSAOs have been identified, a linkage analysis is conducted, which maps each KSAO to relevant tasks. The outcome is a list of critical tasks performed within a job and the KSAOs required to perform each task. This information can then be used to develop selection and certification exams that help to ensure that people have the KSAOs required to successfully perform the tasks within a job.

A primary strength of job analysis is the multistep procedure that is used to gather information about a job. It is important to consider, however, the potential inaccuracies that may occur within the job analysis. For example, two SMEs that perform the same job may have different thoughts on the importance of certain KSAOs, or they may disagree on the frequency in which a particular task is performed. To ensure that the information produced by a job analysis is accurate, it is recommended that multiple SMEs be included, and multiple methods be used for gathering information (Morgeson & Campion, 1997). Job analysis is a complex undertaking, and it produces complex data. This is where the many faceted Rasch model (MFRM) comes into play as an essential tool.

10.3.2 Job analysis and multifaceted measurement modeling

As noted above, job analysis is essential for designing selection systems and certification exams that assess the KSAOs required to effectively perform the tasks within a job. There are, however, concerns regarding the reliability of job analysis frequency and importance ratings. This is largely due to errors in human judgment and genuine differences in opinions across SMEs (Sanchez & Levine, 2012). More simply, there is room for some error in job analysis – and the MFRM has potential to solve some of these concerns.

While it would be impossible to describe all the possible sources of error in job analysis, there are some known issues. Common sources of measurement errors in job analysis ratings can stem from social and cognitive sources such as pressures to portray one's job as important and misperceptions on how frequently certain tasks are performed. Consequentially, these errors influence multiple aspects of job analysis such as the reliability of ratings, interrater agreement, the ability to discriminate between jobs, and the dimensional structure of the data.

To overcome these issues, it has been suggested that multiple raters and methods be used to collect job analysis data (Morgeson & Campion, 1997). Statistical models evaluating multiple sources of variance (e.g., such as generalizability theory) can be used to simultaneously estimate multiple sources of error in job analysis ratings (Morgeson & Campion, 2000). But where statistical approaches do not typically model interval quantities or enable the calibration of a defined unit standard supporting applications across samples and tasks, multifaceted measurement models are not only useful for estimating the variance in job analysis ratings but can meet these other needs as well.

Raters, differing scales (such as criticality and frequency), and job tasks comprise the facets in assessment contexts that can be effectively measured and managed when studies are designed so as to afford the needed comparisons. These facets can also be evaluated using dichotomous or rating scale models prior to using and interpreting the results from data collected with the specific aim of evaluating a multifaceted model. As a result of preliminary work like that, the assessment design and subsequent data analyses can be conducted with increased confidence that the job analysis findings and subsequent test specifications will be relatively free from major source of known bias, and uncertainty (standard errors) will be within the tolerance limits of the decision process.

10.3.3 Modeling the registry data

EMS experts ($N = 398$) participated in a structured rating of the Registry's Task List. This Task List was developed through a process similar to the job analysis as described above. A final step in job analysis for many certification organizations involves subjecting the Task List to this kind of review to directly inform test weights. This is a major step in ensuring every administered test is aligned to expert, representative agreement from within the field.

In this study, respondents rated two scales for each of the 28 tasks that were previously developed by a group of practicing emergency medical professionals. Each scale was a five-point ordinal scale. One scale was devoted to frequency. The other was devoted to criticality (i.e., importance). An informed approach toward these competing ideas was required. That is, a medical skill may be utilized infrequently, but when the skill is needed, it is critical that an emergency medical practitioner has the knowledge to implement the skill. The data from these 398 experts was collected with the specific goal of determining the testing specifications and domain weights relative to considerations such as this.

To that end, specific multifaceted model parameters can be easily converted to testing weights with computational efficiency and ease of interpretation. When there is a direct relation between job analysis results and the weight of tested domains, this method is appropriate and recommended (as is the case within certification organizations). The premise is that a multifaceted modeling approach to developing testing specifications incorporating clear requirements for construct definition, reliability, and parameter separation will outperform more conventional approaches that rely on descriptive statistics or arithmetical metrics (as an example, see Babcock et al. (2020) for a conventional approach).

The multifaceted model specified here includes four parameters estimated within a single frame of reference. The model can be summarized as

$$\mathrm{Ln}(\mathrm{Pnilk}/\mathrm{Pnil}(k-1)) = \mathrm{Bn} - \mathrm{Di} - \mathrm{Si} - \mathrm{Cilk}$$

where Pnilk is the probability P of rater *n* rating task *i*, with a rating *k*, on scale *l*,

and:

Bn concerns the rater severity or leniency of every person providing ratings in the job analysis study;

Di concerns task importance or unimportance;

Si concerns the criticality and frequency of the 28 tasks;

Cilk concerns the category threshold transitions.

10.3.4 Use case summary

This multifaceted measurement model captures the severity parameter associated with the task-rater (i.e., the expert rater, the judge) as well as the overall task importance parameter. But that is not the only aspect of the model. In this application, there is a need to capture two scales (frequency and criticality) using the five-point ordinal rating system within each scale. Every task is calibrated in relation to each category transition threshold. This composite task calibration represents the task importance value used for domain weighting purposes.

Note that the analyst could also incorporate an unrestricted polytomous (partial credit) model instead of the rating scale model so as to estimate a parameter for each individual category threshold. This was not advised by Wang and Stahl (2012) as when that model was estimated the results were exponentially more cumbersome to interpret. In this application, for job analysis, the uniformity of the transition thresholds across items deemed the rating scale model as appropriate.

10.3.5 Use case cautions and procedures

Before looking at the general procedures, some notes are warranted: the position of the parameters *and* the weighting method are important first-line considerations. First, special attention should be paid to the orientation of the Task Importance metric. There are two possible orientations: positive and negative. These are defined as:

1. Negative: The mean Task Importance ratings yielding *lowe*r mean ratings result in higher Task Importance calibrations. Practically speaking, criticality and fre-

quency components that are “harder” to rate highly on the scale are positioned higher on the scale.

2. Positive: The mean Task Importance ratings yielding *higher* mean ratings result in higher Task Importance calibrations. Practically speaking, the calibrations will positively correlate to the mean rating: it is not an inverse relationship like the negative orientation.

Because this analysis scales the measurements in a negative direction (common for educational and psychological measurement), a transformation of the Task Importance score is needed. This is simply accomplished by multiplying the negative position of the Task Importance parameter by −1. The tasks that are easy to rate (high frequency and high criticality) as frequent and critical will then assume the top positions. They will be weighted more heavily when the transformed Task Importance parameter (Di), after moving to the correct positioning, is summed across domains into a total weight. This is important because when these values are summed across the domain, an analyst would want more weight applied to the tasks that are easily frequent and critical.

Second, analysts need to determine if any weighting needs to be applied while the model is being calibrated. Linacre’s guidance in the Facets Software (2023) referred to this as arbitrary weighting. In the case of this modeling technique, arbitrary scale weighting was used to create purposeful metrics. Criticality was assigned 67% of the overall weight of the task importance scale, while frequency was assigned 33% of the overall weight. This weighting makes an adjustment to the obtained raw score during the calibration phase of the modeling technique.

The final Di parameter is then a composite logit measurement of frequency and criticality, which is Task Importance. This Task Importance measure also considers the entire model in its estimation. This helps ensure that critical emergency medical skills would always have slightly more importance than skills that were less critical but perhaps more frequent. Every testing program needs to think through these nuances as a part of their analytical plan. The approach here is what was determined in this unique testing program, but by no means is thus a definitive guideline.

10.4 General procedure

Table 10.1 provides the nine general procedures that are needed to conduct this analysis. Table 10.2 documents the code that can be utilized to achieve a similar analysis. Following this, Tables 10.3–10.5 display the results of this example analysis such that analysts can see the steps involved.

Table 10.1: General procedures.

Step 1	Organize the data to be read into the analysis. See Figure 10.1 for the data schematic.
Step 2	Prepare the code file to fit the four-part model specified above. See the above equation. Special note: consider any weighting that is necessary and incorporate that into the code. This weighting can be based on arbitrary decisions or purposeful concepts that support your measurement plan. The weightings here illustrate a purposeful approach. See Table 10.2 and Figure 10.2 for annotated code guidance.
Step 3	Execute the analysis in using the data schematic and code guidance provided below.
Step 4	Evaluate the data consistency and reliability (e.g., infit, outfit, and point biserial). See Table 10.3.
Step 5	Export and label task importance measures within a spreadsheet application. See Table 10.4 for an example of the basic structure.
Step 6	Transpose the measures so that "easy"-to-rate tasks take on the highest importance value. Multiply the Task Importance parameter by −1 if the facets code is set to place the Task Importance rating in the negative position. See Table 10.4 for an example of the transformations.
Step 7	Convert the Task Importance calibrations (a logit value) into a weighted percentage using two constraints: 1. The sum of the weights must be =100%. 2. The lowest task rating must be fixed to a reasonable starting point, such as 1% or 2%.
Step 8	Sum the transposed percentages within their domains.
Step 9	Review the domain level content weights to ensure they are consistent with the testing goals and representative of the theoretical construct under measurement.

Rater	**Scale**	**Indexing variable**	**Rating responses *(on a five-point scale)***
1	1	1-*N*a	3 3 4 4 5 . . . *N*
.	1	1-*N*a	3 3 1 4 5 . . . *N*
N	1	1-*N*a	2 2 3 3 4 . . . *N*
1	2	1-*N*a	1 1 3 3 5 . . . *N*
.	2	1-*N*a	1 3 2 4 5 . . . *N*
N	2	1-*N*a	1 1 3 2 4 . . . *N*

Figure 10.1: Example data schematic for four-part multifaceted model.

To fit the four-part model, the above data schematic is utilized. In the above example, a five-point rating scale is used for illustration purposes. In the study data presented here, there were 28 variables under the "Rating Responses" heading. One column for each task was evaluated. A unique rater number (the Rater ID) is entered on the row, followed by the scale that is being rated (e.g., frequency, importance, and criticality).

This rating scale is assigned a unique number (for two scales, there would be two values: 1 and 2).

An indexing variable is placed in the matrix. The letter *N* is the total number of columns of responses. The letter “a” is present only for convenience: when this data is moved back into a spreadsheet application, 1 – *N* (whatever number *N* is) will not convert to a date data type with the attachment of “a” to the field. This index signals to the many-facet software (Linacre, 2023) that *N* number responses are going to be read into the matrix, assigned to the specific rater, and assigned to the specific scale. This continues until the end of the responses on scale 1.

Following the responses associated with scale 1, a new matrix of data is placed immediately below all scale 1 responses. The unique Rater ID starts over again for Rater ID number 1, and the responses for scale number 2 are provided. Readers will note that a new scale identifier is used to signal a new scale. Similarly, to scale 1, the indexing variable is present in scale 2, and the responses are read into the many-facet estimation.

The general data conventions include:

1. every rater has a unique Rater ID;
2. a scale has a single unique numerical ID indicating which scale it is;
3. an indexing variable that indicates what responses are going to be read in and assigned to the rater and the scale; and
4. response strings that are the actual obtained ratings on the scale (criticality and frequency, in this case from the expert raters).

Figure 10.2 provides the core code for fitting this four-part model. Table 10.2 explains the code in more detail. For novice analysts using Linacre’s (2023) software, the model = command line is the most important to become familiar with. That line of code is where the model is specified. See Linacre (2023) for help documentation to learn what choices and options are available:

10.5 Evaluating the results

Probabilistic measurement models specify the requirements for estimating interval quantities, and so provide a powerful framework for evaluating the results of activities intended to ordinally rank or score performances. Common methods of evaluating modeling results include information-weighted (infit) and outlier-sensitive (outfit) statistics, point-biserial correlations, and principal component analysis. The results from the fit of the four-part model are described in Table 10.3. The data showed satisfactory fit to the model, demonstrated that raters consistently related task criticality and frequency to the latent construct (task importance), and preliminary principal component analysis results indicate the construct (task importance) to be usefully unidimensional.

```
Title = Example Job Analysis Data
convergence = 0.1;
unexpected = 3.0;
arrange = M;
facets = 3;
noncenter = 1;
negative = 1,2,3;
Inter-rater = 1;
pt-biserial= measure;
usort = 1,2,3;
Models =
?,#,?,R5;
*
Labels =
1,Raters,
1-398
*
2, Scale,
1=Critical,,,2.0
2=Freq,,,1.0
*
3, Measures,
1=nD1T1F
*
```

Figure 10.2: Code layout for many facets control file in the Facets software.

Table 10.2: Important code descriptions.

Code	Description
Arrange = M	Output tables will be placed in ascending order.
Facets = 3	There are three facets: Raters, Scales, and Measures. *However, there are two rating scales. So, four parts are ultimately calibrated.*
Noncenter = 1	The floating facet is the rater. The rest of the facets will be scaled to center on zero. They will not float.
Inter-rater = 1	There is an inter-rater agreement that is expected. The rater facet, #1, is designated to be where the inter-rater agreement is calculated.
Models = ?,#,?,5	Raters are assigned to a common rating scale; each scale is allowed to have its own structure; measurements are estimated in relation to a common rating scale; there are five valid rating values (0 to 4, directly matches the Likert scale format).

Table 10.2 (continued)

Code	Description
Labels = 1, Raters, 1–398 * 2, Scale, 1 = Critical,,,2.0 2 = Freq,,,1.0 * 3, Measures, 1 = The name of the first measure, 2 = The name of the second measure, *N* . . . = The name of the nth measure.	The Labels command here shows Raters as the first facet. There are 398 of them and each is numbered uniquely in the 1–398 range. Scales are the second facet. There are two of them, making this a four-component model. The Criticality Scale is weighted 2× higher than the Frequency Scale; algebraically equivalent to 67% of value going to Criticality Scale while 33% weight goes toward frequency. Weights go in the third comma position. Measures are the third facet. Each measure has a name.

Table 10.3: Summary of model performance.

Component	Desired and optimal values	Obtained values from present study
Rater-level statistics		
Rater mean infit	0.50 through 1.50	1.10
Rater mean outfit	0.50 through 1.50	1.15
Rater mean correlation	>0.30	0.32
Rater reliability	0.8	0.87
Scale-level statistics		
Scale mean infit	0.50 through 1.50	1.1
Scale mean outfit	0.50 through 1.50	1.31
Criticality mean correlation	>0.30	0.56
Frequency mean correlation	>0.30	0.37
Overall scale reliability	0.8	0.96
Task importance measure statistics		
Task importance mean infit	0.50 through 1.5	1.04
Task importance mean outfit	0.50 through 1.5	1.15
Task importance mean correlation	>0.30	0.43
Task importance reliability	0.8	0.98
Explained variance	Overall explained variance in this single model: 71%	

10.6 Working with the results

After the above code is executed and the results are evaluated, Table 10.4 can be constructed from the program output. This table takes the label for each task from the Task List, and places it next to the mean rating provided from the raters' responses on the frequency and criticality scales. It then displays the measurement associated with the task. Remember: each task is associated with a specific testing domain and each task will end up with a task importance calibration that is estimated by the raters' mean scores on criticality and frequency. This is then adjusted for the weight of the scale (67% for criticality and 33% for frequency) to produce the composite Di parameter. The transformed Di parameters (Di × −1) are then summed within their domain to arrive at the result. As readers will see, this parameter gives tremendous insight into the importance of the task and in the domain based on the raters, the ratings, and the scales.

Keep in mind, and this cannot be emphasized enough: the task importance calibrations (Di) are negatively oriented and needs to be placed into a transformed position. This is accomplished by multiplying the calibration by −1. It is important to do this if the calibration data was estimated with the negative position. Think about this in a practical way: Task 5 on Dimension 4 is the hardest task to rate as being critical or frequent. As a result, this task needs the lowest relative placement of weighting. Each testing program will have to determine where to start the weighting process. In this example, the transformed weight originates at 1%; and 2% and 3% origins were also estimated to help make final determinations. However, in Table 10.4, the lowest Di estimate is directly converted to a 1% task weight with the described constraints.

Table 10.4: Labeled and transformed task importance calibrations.

Dimension no. (D) and task no. (T)	Mean ratings	Task importance calibrations	Transformed calibrations	Weight: 1% origin
D4 T5: Manage interventions to special populations	2.66	0.56	−0.56	1.00
D5 T4	2.66	0.56	−0.56	1.00
D5 T3	2.69	0.53	−0.53	1.14
D1 T3	2.74	0.47	−0.47	1.41
D1 T1	2.76	0.46	−0.46	1.46
D4 T7	2.83	0.38	−0.38	1.82
D2 T1	2.85	0.36	−0.36	1.92
D4 T3	2.86	0.35	−0.35	1.96

Table 10.4 (continued)

Dimension no. (D) and task no. (T)	Mean ratings	Task importance calibrations	Transformed calibrations	Weight: 1% origin
D1 T6	2.88	0.33	−0.33	2.05
D4 T4	2.93	0.27	−0.27	2.33
D4 T6	3.02	0.16	−0.16	2.83
D1 T4	3.05	0.12	−0.12	3.02
D3 T2	3.06	0.11	−0.11	3.06
D3 T1	3.07	0.1	−0.1	3.11
D1 T5	3.08	0.08	−0.08	3.20
D1 T2: Protect self, other responders, patient, public, and the emergency scene from existing and potential hazards	3.15	0	0	3.57
D5 T2	3.22	−0.1	0.1	4.02
D2 T2	3.27	−0.18	0.18	4.39
D2 T3	3.31	−0.23	0.23	4.62
D2 T8	3.34	−0.28	0.28	4.85
D4 T2	3.36	−0.31	0.31	4.99
D5 T1	3.38	−0.34	0.34	5.12
D4 T1	3.43	−0.44	0.44	5.58
D2 T5	3.45	−0.47	0.47	5.72
D2 T9	3.49	−0.54	0.54	6.04
D2 T7	3.5	−0.57	0.57	6.18
D2 T6	3.51	−0.59	0.59	6.27
D2 T4: Assess the patients’ airway	3.61	−0.82	0.82	7.32
Sum				100%

Note: This table displays the exact results from the Emergency Medical Technician Task Rating Data. Mean ratings are transformed into a calibration, per the model specifications.

A linear solution, with two constraints, is required to convert the transformed calibration into a relative weight. Most psychometricians are familiar with this process as linear transformations are commonly employed. For those interested, this can be accomplished in Microsoft Excel via the Excel Solver package (Excel Solver, 2023):

1. Constraint #1: The summation of the weights cannot exceed 100.
2. Constraint #2: The first starting value must be fixed to the predetermined origin. In this example, the origin was set to 1%. Every analyst and program should decide on an origin that makes sense for their own unique context.

Once each task is assigned a weight, those weights can be summed and used to inform testing specifications. In this example, we see there are five domains for every test administration. Based on the results of the MFRM, the lowest dimension should represent 6% of the weight while the highest dimension should represent 47% of the weight. Of course, this number will vary depending on where an analyst and a testing team start their origin. In this example, the origin started at 1%. When the origin starts at a higher percentage, these weighted sums move closer toward one another. The result is that each tested domain arrives at reasonable content weight for building blueprints and specifications. Examination content committees can then use these defensible weights as a starting point for determining testing weights (see Table 10.5).

Table 10.5: Summation of the tasks by domain for test content weighting.

Domain	Percentage of representation on test forms: using a 1% origin	Final adopted percentage
Domain 1: Scene size-up and safety	14.7%	15–19%
Domain 2: Primary assessment	47.3%	39–43%
Domain 3: Secondary assessment	6.2%	5–9%
Domain 4: Patient treatment and transport	20.5%	20–24%
Domain 5: Operations	11.3%	10–14%

10.7 Discussion and summary

This chapter illustrates a multifaceted measurement modeling approach to designing, analyzing, and reporting job analysis data. Job analyses are used to directly inform testing specifications and testing blueprints for certification organizations. It is imperative that certification organizations have defensible methods for job analysis data because job analysis determines content weights on a test. The method here extends the work of Wang and Stahl (2012) by providing more explicit instructions on the procedures and by sharing an example with the code.

In this example, 398 experts in the EMS field rated 28 tasks for criticality (importance) and frequency. Those ratings were then modeled in a four-part process in which task importance ratings were captured in the model parameter, Di. That parameter was then weighted and summed across tested domains. The testing specification was modeled directly from that parameter. In this example, when the lowest task importance score is set to 1%, the range of domain weights is between 6% and 47%.

The range changes based on the location of the origin. It may be important to evaluate multiple origins as the data are analyzed and interpreted.

Analysts need to think carefully about facet positioning. The illustrated code here places all facets in a negative position. This is common in educational and psychological measurement but is *not* common in scales used in healthcare settings. To that end, the analyst will need a method, like the one shared here, to evaluate how easy-to-rate tasks take on higher weights in the final weighting solution.

Rarely does a single psychometrician have the final say on test specifications. Final content weights and testing specifications are almost always determined by a group of decision-makers, such as a standards and examinations committee. In this case, a content-weighted analysis provided objectively reproducible and explained results supporting a consensus decision process. When committees are working through complex rules for test design, test specifications, and making decisions about format and delivery, setting a high bar for comparability by modeling a fair basis for the inferences to be made, as is illustrated here, the credibility and defensibility of committee recommendations and decisions are significantly enhanced.

Finally, we encourage readers to consider and utilize multifaceted measurement modeling when building testing specifications. Raters, scales, and other components can be located on a single scale enabling comparisons in a common frame of reference. This single scale creates a common understanding demonstrably based in evidence, promotes defensible decisions, and enhances the interpretability of job analysis findings by including components beyond the two facets of person and item location. While there are many ways to implement job analyses, scaling the results on a common interval unit of measurement offers the objectivity needed for scientifically defensible comparisons.

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*, 449–460.

Andrich, D. (1988). *Sage University Paper Series on Quantitative Applications in the Social Sciences. Vol. series no. 07–068: Rasch models for measurement*. Sage Publications.

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social, and health sciences*. Springer.

Babcock, B., Risk, N., & Wyse, A. (2020). On the superior statistical properties of frequency scales in job analyses. *Educational Measurement: Issues and Job*, *39*(2), 85–95.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olson-Buchanan (Eds.). *Innovations in computerized assessment* (pp. 67–91). Lawrence Erlbaum Associates, Inc., Publishers.

Bond, T., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences*. 4th ed. New York: Routledge.

Cano, S., Vosk, T., Pendrill, L., & Stenner, A. J. (2016). On trial: The compatibility of measurement in the physical and social sciences. *Journal of Physics: Conference Series*, *772*, 012025. doi: 10.1088/1742-6596/772/1/012025

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Statistics for Social and Behavioral Sciences. Springer-Verlag.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. 2nd ed. Peter Lang

Engelhard, G., Jr. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement: Interdisciplinary Research and Perspectives*, *6*(3), 155–189.

Excel Solver (2023). Excel solver tutorial. *Front Line Solvers*. https://www.solver.com/solver-tutorial-using-solver

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Fisher, W. P., Jr. (2003a). The mathematical metaphysics of measurement and metrology: Towards meaningful quantification in the human sciences. In A. Morales (Ed.). *Renascent pragmatism: Studies in law and social science* (pp. 118–153). Ashgate Publishing Co.

Fisher, W. P., Jr. (2003b). Mathematics, measurement, metaphor, metaphysics: Parts I & II. *Theory & Psychology*, *13*(6), 753–828.

Fisher, W. P., Jr. (2004). Meaning and method in the social sciences. *Human Studies: A Journal for Philosophy and the Social Sciences*, *27*(4), 429–454.

Grosse, M. E., & Wright, B. D. (1986). Setting, evaluating, and maintaining certification standards with the Rasch model. *Evaluation & the Health Professions*, *9*(3), 267–285.

Kelley, P. R., & Schumacher, C. F. (1984). The Rasch model: Its use by the National Board of Medical Examiners. *Evaluation & the Health Professions*, *7*(4), 443–454.

Li, G., Pan, Y., & Wang, W. (2021). Using generalizability theory and many-facet Rasch model to evaluate in-basket tests for managerial positions. *Frontiers in Psychology*, *12*, 660553.

Linacre, J. M. (1994). *Many facet Rasch measurement*. Chicago: IL: University of Chicago Press.

Linacre, J. M. (1995). Paired comparisons with ties: Bradley-Terry and Rasch. *Rasch Measurement Transactions*, *9*(2), 425. http://www.rasch.org/rmt/rmt92d.htm

Linacre, J. M. (2000). Almost the Zermelo model? *Rasch Measurement Transactions*, *14*(2), 754. http://www.rasch.org/rmt/rmt142k.htm

Linacre, J. M. (2023). *Facets computer program for many-facet Rasch measurement, version 3.84.0*. Beaverton, Oregon: Winsteps.com.

Linacre, J. M., Engelhard, G., Tatum, D. S., & Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research*, *21*(6), 569–577.

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement*, *51*, 315–327. http://www.sciencedirect.com/science/article/pii/S0263224114000645

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement*. 2nd ed. Springer Series in Measurement Science and Technology. Springer. https://link.springer.com/book/10.1007/978-3-031-22448-5

Melin, J., Cano, S., & Pendrill, L. (2021). The role of entropy in construct specification equations (CSE) to improve the validity of memory tests. *Entropy*, *23*(2), 212.

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology*, *82*, 627–655. https://doi.org/10.1037/0021-9010.82.5.627

Morgeson, F. P., & Campion, M. A. (2000). Accuracy in job analysis: Toward an inference-based model. *Journal of Organizational Behavior*, *21*, 819–827. https://doi.org/10.1002/1099-1379(200011)21:7<819::AID-JOB29>3.0.CO;2-I

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*, 371–389.

Newby, V. A., Conner, G. R., Grant, C. P., & Bunderson, C. V. (2009). The Rasch model and additive conjoint measurement. *Journal of Applied Measurement*, *10*(4), 348–354.

Pendrill, L. R. (2014). Man as a measurement instrument [Special Feature]. *NCSLi Measure: The Journal of Measurement Science*, *9*(4), 22–33. http://www.tandfonline.com/doi/abs/10.1080/19315775.2014.11721702

Pendrill, L. R. (2019). *Quality assured measurement: Unification across social and physical sciences*. Springer Series in Measurement Science and Technology. Springer. https://link.springer.com/book/10.1007/978-3-030-28695-8

Pendrill, L., & Fisher, W. P., Jr. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement*, *71*, 46–55. http://dx.doi.org/10.1016/j.measurement.2015.04.010

Sanchez, J. I., & Levine, E. L. (2012). The rise and fall of job analysis and the future of work analysis. *Annual Review of Psychology*, *63*, 397–425. https://doi.org/10.1146/annurev-psych-120710-100401

Smith, R. M., Julian, E., Lunz, M., Stahl, J., Schulz, M., & Wright, B. D. (1994). Applications of conjoint measurement in admission and professional certification programs. *International Journal of Educational Research*, *21*(6), 653–664.

Stenner, A. J., Fisher, W. P., Jr., Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology: Quantitative Psychology and Measurement*, *4*(536), 1–14. doi: 10.3389/fpsyg.2013.00536

Tavakol, M., & Pinner, G. (2019). Using the Many-Facet Rasch Model to analyse and evaluate the quality of objective structured clinical examination: A non-experimental cross-sectional design. *BMJ Open*, *9*(9), e029208.

Wang, N., & Stahl, J. (2012). Obtaining content weights for test specification from job analysis task surveys: An application of the many-facets Rasch model. *International Journal of Testing*, *12*(4), 299–320. https://doi.org/10.1080/15305058.2011.639472

Warner, D. O., Isaak, R. S., Peterson-Layne, C., Lien, C. A., Sun, H., Menzies, A. O., . . . Harman, A. E. (2020). Development of an objective structured clinical examination as a component of assessment for initial board certification in anesthesiology. *Anesthesia & Analgesia*, *130*(1), 258–264.

Wilson, M. R. (2023). *Constructing measures: An item response modeling approach*. 2nd ed. Routledge

Wind, S., & Hua, C. (2022). *Rasch measurement theory analysis in R*. CRC Press.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*(4), 33–45, 52. [http://www.rasch.org/memo62.htm]. https://doi.org/10.1111/j.1745-3992.1997.tb00606.x

Simon Karlsson, Hanna Svensson, Jacob Wisén, and Jeanette Melin

11 A metrological approach to social sustainability metrics in municipalities

Abstract: This chapter presents a metrological approach for measuring social sustainability aspects in municipalities: the aim is explicitly formulated in support of a much-needed extension of the SI units. The approach builds on rigorous conceptual definitions of the attributes to be measured, includes a focus on usability, and extends the recently developed model "man as measurement instrument" – including the application of probabilistic conjoint measurement theory – into social sustainability metrics in municipalities. In Section 11.2, the collaborative process of defining the constructs to be measured is described. This is followed by Section 11.3, which defines our measurement system and measurement model where the "municipality as measurement instrument" is introduced together with the unique metrological properties obtained. In Section 11.4, a testing procedure for theoretical expectations is presented, and we then further illustrate this approach based on a sustainability aspect – quality of life – using data from the Swedish open database, Kolada, in Section 11.5. Finally, this chapter concludes with reflections, learning, and proposals for further methodological advancements. We encourage a collaborative process between end-users, social sustainability experts, and metrologists – for developing, testing, evaluating, and revising fit-for-purpose metrics and, thus, enable more well-informed decisions in developing meaningful and effective policies.

Keywords: construct definition, measurement system analysis, measurand restitution, social sustainability, municipality policy

Acknowledgments: The authors would like to express their gratitude to the staff of City of Helsingborgs for their time, support, and willingness to cooperate. The authors would also like to give a particular thank to senior support from Leslie Pendrill, RISE/Research Institutes of Sweden, William P. Fisher, Jr., Living Capital Metrics LLC, Sausalito CA USA/BEAR Center, Grad. School Education, UC Berkeley, RISE, Research Institutes of Sweden.

Part of this work was supported by the RISE Platform Center for Category-Based Measures and RISE strategic initiative Measurement for Sustainable Transformation.

Simon Karlsson, **Hanna Svensson**, **Jacob Wisén,** RISE Research Institutes of Sweden, Division of Safety and Transport, Measurement Science and Technology Unit, Gothenburg, Sweden
Jeanette Melin, RISE Research Institutes of Sweden, Division of Safety and Transport, Measurement Science and Technology Unit, Gothenburg, Sweden; Swedish Defence University, Department of Leadership, Demand and Control, Karlstad, Sweden

https://doi.org/10.1515/9783111036496-011

11.1 Introduction

Social sustainability is an increasingly prioritized topic within public, private, and civic organizations and in national, regional, and local level governments worldwide. However, aspects of social sustainability lack common definitions and objective measures, which in turn risks biases and confusion, rather than an understanding of the characteristics of sustainability itself (Fisher & Pendrill, 2019). We argue that it is time for a "social sustainability metric system" where countries, regions, and municipalities can benchmark their current position relative to where they have been, where they want to go, and what to do next, which would improve the work towards sustainable societies (Svensson et al., 2022).

Measuring is never an end itself. It is a way of gaining knowledge about the world to make well-informed decisions. For example, in the realm of social sustainability, analysts may want to identify the current state of a population's health, wellbeing, social inclusion, or participation in different regions to prioritize where intervention is most needed. To perform this task successfully, the information underlying the decision must be correct, which in turn calls for a quality-assured measurement system. This system includes specifications for quality properties, such as validity and reliability, which also should be coordinated with end-users. If the measurement system does not match their needs, it may not be used and if it is not valid and reliable it will result in incorrect conclusions.

Despite decades of evidence supporting the feasibility, viability, and desirability of metrological traceability in social sciences (cf. Wright, 1997; Cano et al., 2019; Pendrill & Fisher, 2015; Pendrill, 2019; Wilson, 2013; Mari & Wilson, 2014), a metrological approach for constructing reasonable and meaningful measurements of social sustainability aspects in municipalities has not yet been demonstrated. Currently, typical evaluations of social sustainability aspects are based on sets of indicators compiled and summarized via manual analysis. Nevertheless, this procedure is time-consuming due to the often substantial number of indicators, and it is challenging to obtain meaningful information about social sustainability and its management due to its multidimensional nature (Svensson et al., 2022). Moreover, there is an urgent need to replace incomparable counted fractions, ordinal ratings, and counts, with quality-assured social sustainability metrics (Fisher, Melin & Möller, 2021; Fisher, 2020a/b, 2022) to develop more meaningful and effective municipal policies and sustainability programs.

Given the current needs and challenges, this chapter presents principles for a metrological approach to construct reasonable and meaningful measurements of social sustainability aspects in municipalities. The approach emphasizes the end-user, builds on rigorous conceptual definitions of the attributes to be measured, and extends the recently developed model 'man as measurement instrument' into social sustainability metrics in municipalities.

11.2 A collaborative process for construct definitions

To construct reasonable and meaningful measurements, we propose an iterative process that incorporates both qualitative and quantitative perspectives. Our approach builds on Mark Wilson's (2005) construct modeling as well as other advancements in test theory and measurement science (cf. Pendrill, 2019; Boone, 2016; Hobart & Cano, 2009; Andrich & Marais, 2019; Melin & Pendrill, 2023; Pendrill & Melin, 2023; Salzberger, 2019). The initial steps focus on understanding the purpose of the measurement by aligning the qualitative definition of the construct to be measured with user needs, creating a construct map, and then selecting or designing items. Later steps highlight the data management and a description of the measurement system being used, which will follow in Sections 11.3 and 11.4.

An important feature of the recommended procedure is that it encourages a collaborative process between end-users, social sustainability experts, and metrologist, which is essential for developing user-friendly measurement tools based on domain knowledge and daily experience. In addition, metrologists specializing in the social sciences possess in-depth knowledge in measurement science, construct modeling, and item generation, which sustainability experts and end-users usually do not possess. To the best of our knowledge, this form of expertise is absent in policymaking processes, where there is a widespread use of various metrics and indexes with varying metrological quality.

Initially, the most fundamental question to ask is *what* construct the end-user is interested in measuring. The answer to this question should set the boundaries for upcoming activities, as the main objective is to create favorable conditions for measuring the construct relevant to the end-user. For example, QoL is a construct of interest in many sustainability policy programs (Mohamed & El-Walid, 2019; Fors, 2012). Thus, if well-informed decisions about quality of life (QoL) are to be made consistently throughout the system, we must understand what QoL is and what potential subcomponents it consists of.

There are multiple ways to obtain a deeper qualitative understanding of constructs. An initial step is to examine the existing literature on the topic and search for previously developed conceptual models (U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, & U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health, 2006). If such literature is missing, a good start is to consult experts in the field or conduct focus groups. Importantly, questions about content validity should be addressed before proceeding with other measurement properties (Morel & Cano, 2017).

The construct to be measured may be multidimensional. If this is the case, one must ask at what level the decisions are to be made; for example, the overall level of social sustainability or a specific sub-aspect of the construct. To avoid future problems regarding dimensionality, it is recommended to consider the construct that we aim to measure as a unidimensional latent variable. As put by Wilson, we assume that *the construct we wish to measure has a particularly simple form – it extends from one extreme to another, from high to low; small to large, positive to negative, or strong to weak* (Wilson, 2005 p. 6). This is accomplished by sketching out the definition of the construct to be measured in a construct map, where the focus is to broadly describe the characteristics of being located at different points of the continuum. If the construct of interest is multidimensional, which social sustainability is regarded to be, it is recommended to handle each aspect separately, formulating a construct map for each aspect (Wilson, 2005).

For instance, a common inquiry among municipalities is to gain knowledge about their ability to provide social sustainability to its citizens. To gain this knowledge, a measurement system for the construct of interest is required, where the first step would be to write a description of what would characterize a municipality with a low sophistication level, and what it would mean for a municipality to have a progression in sophistication level on the same latent dimension. As will be further described in Section 11.3, the measurement system will consist of more challenging indicators as well as municipalities with lower to higher ability of the construct of interest. If indicators can be designed or selected along the continuum from easier to more challenging tasks (mapped by a pattern of consistent changes in the response likelihoods), this would enable a more fit-for-purpose measure of the municipalities' abilities (Fisher, Melin & Möller, 2021). Accordingly, a comprehensive set of indicators is desired, with different items capturing the various levels of the underlying latent variable that is being measured. When individuals are under scrutiny, real-world observations are derived from people answering surveys, taking tests, or performing tasks to calibrate the measurement system. As such, it is possible to develop items and measurement tools by letting respondents answer questions and, in that way, gather data (Wilson, 2005; Boone, 2016; Hobart & Cano, 2009; Andrich & Marais, 2019). When municipalities are under scrutiny the real-world observations are harder to get and instead, aggregated data from multiple municipalities and several data sources could be required to calibrate the measurement system. This requirement makes it a complex and challenging task to construct municipality-level indicators from scratch, implying that we often have to rely on existing indicators gathered in multiple geographical areas.

To gather relevant indicators to measure social sustainability aspects at the municipal level, we recommend the following procedure:

i. Determine the geographical level of the measurement. If the aim is to measure a construct at the municipality level, the indicators must be available at the municipality level.

ii. Map relevant data sources. The mapping should consider both constructs (i.e., does the data source provide data related to the domain of interest?) and the desired geographical data level (i.e., does the data source provide data at the desired geographical level?).

iii. Map relevant indicators. At this point, we recommend returning to the conceptual theory and constructing a map to inform the selection of indicators.

The screening and removal of non-functioning items will be handled in a subsequent phase. To ensure a successful calibration of our measurement system later, it is crucial to incorporate a range of abilities for municipalities and a variation in the difficulty levels of the items.

When gathered the relevant indicators, it is suggested to formulate an ordinal item theory (Melin, Fisher & Pendrill, 2021) to obtain an overview of all the indicators related to the construct identified during the data screening process. The main objective is to create the foundations for a successful match between the distribution of the study sample (in this case, municipalities) and the indicators (Hobart & Cano, 2009), and, in turn, lay the best foundation possible for meaningful measurements of sustainability aspects in municipalities.

11.3 Measurement system analysis and measurand restitution

Another critical step in all kinds of measurements is to make a description as complete and correct as possible of the actual measurement system used. In a typical measurement system, the measurement information is transmitted from the measurement *object* to an *instrument* and then transmitted to an *operator*. These three components are the main elements of the measurement system; however, the measurement *method* and *environment* can also influence the measurement system when determining the overall measurement quality (E11 Committee). This is a well-established approach in traditional metrology (Bentley, 2005; Loftus & Giudice, 2014; E11 Committee) but has not yet been fully recognized within the human and social sciences.

Pendrill argued that *drawing simple analogies between 'instruments' in the social sciences questionnaires, ability tests, etc. and engineering instruments such as thermometers does not go far enough* (Pendrill, 2018, p. 1). Thus, a decade ago, the model 'man as measurement instrument' was introduced (Berglund et al., 2012), which is a measurement system approach adapted for the social sciences (Pendrill, 2014). Building on traditional metrology, two key points justify this approach:

i. A mass standard is the primary choice of metrological reference, as it is typically robust and simple, in preference to choosing a sensitive and complex weighing instrument (Pendrill, 2019).
ii. In contrast to the robustness and simplicity of questionnaire items, humans are more complex and sensitive to the environment and method; consequently, they are less suitable for use as metrological references (Pendrill, 2021; Melin, 2021).

A measurement system approach to measurements in the social sciences implies that the measurement *object*, such as a questionnaire item with a certain level of difficulty, provides a stimulus to the *instrument*, that is, the person who provides a response, such as a pass or fail, to an *operator* (Pendrill, 2014; Pendrill, 2019) (Figure 11.1). Thus, the response received by the *operator* is a combination of the difficulty of the item and the ability of the person: a person with high ability is more likely to respond positively to a difficult item than a person with lower ability, and an easy item is more likely to be endorsed by more persons than a more difficult item (this is further exemplified in Section 11.5 for social sustainability metrics, as illustrated in Figure 11.2). Consequently, the response is characterized by having no numerical meaning, and it can only be used to indicate order (Turetsky & Bashkansky, 2022; Wright & Linacre, 1989). This response is remarkably similar to what is typically observed in today's evaluations of social sustainability based on sets of compiled and summarized indicators. Thus, we would argue for extending the model 'man as measurement instrument' into social sustainability metrics in municipalities where a corresponding measurement system for social sustainability metrics in municipalities will imply that the measurement *object*, such as the indicator with a certain level of difficulty, δ_i, provides a stimulus to the *instrument*, that is, the municipality, who responds depending on its ability, θ_n.

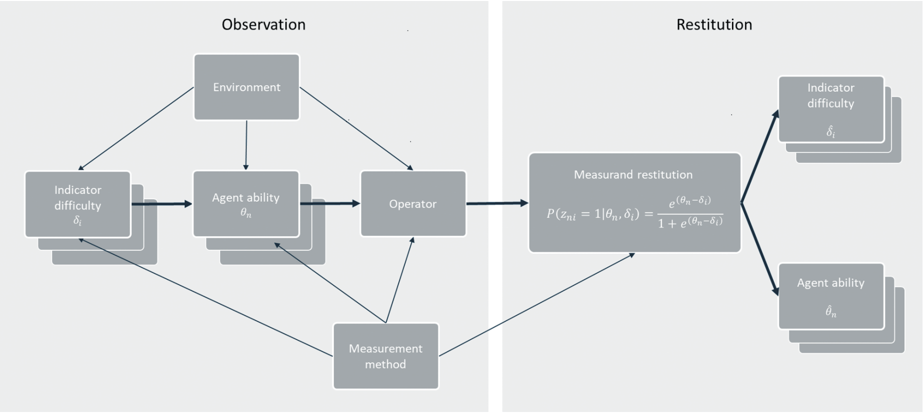

Observation
Environment
Indicator difficulty δ_i
Agent ability θ_n
Operator
Measurement method
Restitution
Measurand restitution
$P(z_{ni} = 1|\theta_n, \delta_i) = \frac{e^{(\theta_n - \delta_i)}}{1 + e^{(\theta_n - \delta_i)}}$
Indicator difficulty $\hat{\delta}_i$
Agent ability $\hat{\theta}_n$

Figure 11.1: A measurement system approach to measurements in the social sciences showing the processes in the observation phase and the restitution phase, respectively.

In line with traditional metrology, both when humans and municipalities act as instruments, the observed responses require a *measurand restitution* (Pendrill, 2018; Rossi, 2014) as a key component to extend the SI to the human and social sciences. Ordinal ratings and counts can then be restituted into linear and separate measures for the two coupling attributes, that is, the object and instrument attributes, on a conjoint scale where a specific kind probabilistic measurement model is particularly suitable. These models belonging to what has been shown to be a general class of equivalent models (Andrich, 1988, p. 43; Linacre, 1995, 2000a/b; Newby et al., 2009) independently devised by a number of mathematical innovators working in the early and mid-twentieth century, including Bradley and Terry (1952), Luce (1959), Luce and Tukey (1964), and Rasch (1960), in particular, but also others, such as Levy and Kolmogorov (Wright, 1997). Significant conceptual correspondences with the works of Loevinger (1965), Thurstone (1959), and Guttman (1950) have also been noted (Andrich, 1978b; Engelhard, 2008). In addition, given their general correspondence with metrological conceptions of measurement (Mari & Wilson, 2014; Mari et al., 2023; Pendrill, 2014, 2019), it seems past time to cease foregrounding the name of just one of several important innovators, as each contributor displayed highly creative originality. A general form for these models is

$$P(z_{ni} = 1 | \theta_n, \delta_i) = \frac{e^{(\theta_n - \delta_i)}}{1 + e^{(\theta_n - \delta_i)}} \tag{11.1}$$

where the probability P of a response is a function of the difference between θ_n and δ_i. The ability of agent n is denoted θ_n, and the difficulty of indicator i is denoted δ_i. In psychometrics, the probability P is often set in relation to success on a task or item in a test or questionnaire. In the case of social sustainability metrics in municipalities, this implies that the indicator difficulty, δ_i, gives a measurement value on how difficult the indicator is to achieve, and that the municipality ability, θ_n, provides a measure of how well the municipality supports citizens in various social sustainability aspects (Svensson et al., 2022).

Many of the indicators used in municipalities' reports of social sustainability aspects are so-called counted fractions, limiting the comparability across the continuum as the steps are not equally distributed (Pendrill, 2019; Pendrill & Melin, 2019). This corresponds much to the ordinality issues typically seen with responses to rating scales, as 0–100% can be seen as a polytomous scale with many categories. However, a polytomous scale with more than seven response categories seldom worked well (Simms et al., 2019; Toland et al., 2020; Kersten, White & Tennant, 2014). Therefore, an option is to categorize the percentages into a less nuanced polytomous scale, similar to the procedure for collapsing disordered thresholds (Andrich, 2013), to be able to combine indicators into reasonable and meaningful measurements of sustainability aspects in municipalities (Svensson et al., 2022; Fisher, Melin & Möller, 2021; Fisher, 2020a).

In the case of social sustainability metrics for municipalities, one has to specify how the percentage should be converted into ordinal scores (Fisher, 2020b) based on a set of rules (i.e., pre-specified intervals or cut-off values). In the simplest case, the same intervals or thresholds could be used for all indicators included in the measurement of a construct (i.e., the interval 95–100% is converted into the highest possible score for all items included in the scale). However, because some indicators may be more difficult than others, applying the same rule for categorization may lead to a loss of information. For instance, if the specified intervals for some items are out of reach for all or most municipalities. At the same time, we do not recommend applying a rule for categorization where the cut-off values are solely based on the relative difficulty level of the items. If the categorization rules are set based on the performance of the study sample, such as percentiles of percentages or any other distributional statistics, all items may be found equally difficult to obtain in the subsequent analysis, resulting in a loss of information.

To deal with the issues concerning categorization, we recommend that the cut-off values be set in coordination with experts and/or end-users, where established levels of high and low can often be found in established agreements, norms, and protocols The main objective of this process is to formulate a common grading system that specifies what constitutes *high* or *low* grades This grading should result in a transformation table, where it is clear how different intervals of raw scores should be converted into ordinal scores to be used in measurement and restitution.

In sum, measurand restitution enables the separation of measures for indicator and municipality attributes, which is necessary for providing quality assurance in terms of metrological traceability and measurement uncertainty. To ensure metrological traceability for social sustainability in municipalities, the indicators are our primary choice of metrological standards, in much the same way as in traditional metrology. Although this approach to measurement modeling has been applied in our previous work and similar contexts (Svensson et al., 2022; Fisher, Melin & Möller, 2021; Fisher, 2020b), we argue that a proper description of the measurement system, as presented above, has been lacking. This component is a critical component for a metrological approach to human and social sciences in general when extending the SI (as can be further read in accompanying chapters (Melin, 2023; Pendrill, 2023).

11.4 Theory testing and validation

As mentioned in Section 11.2, the construction of reasonable and meaningful measurements requires an iterative process consisting of multiple steps. In the previous sections, the main focus was on harnessing our theoretical knowledge of the foundations for constructing reasonable and meaningful measurements of social sustainability aspects. This section includes the steps for empirical testing of whether theoretical expectations find support in real-world data.

In short, theory testing and validation imply that the outline of the construct map and ordinal theory is examined against the hierarchical output from the measurement and restitution. Thus, it is necessary to first assess whether the basic principles of measurement quality assurance are satisfied. It has recently been proposed to evaluate the following criteria (Johansson et al., 2023):

i. **Unidimensionality**: Do indicators represent one latent trait without strongly correlated item residuals?
ii. **Response categories**: Do municipalities' overall positioning on the scale reliably predict the ordinal score they obtain for each indicator?
iii. **Invariance**: Do the indicators work in the same way for relevant subgroups (e.g., over time and between small and large municipalities)?
iv. **Targeting**: Is the location of the indicator threshold matched with the location of the municipalities and does not show strong ceiling or floor effects or gaps?
v. **Reliability**: Is the reliability sufficient for the expected properties of the target population and intended use?

If these basic criteria are satisfied, the hierarchical order of the indicators can be compared with the expected construct map and the ordinal theory developed in the previous steps. If there is a good match between the empirical hierarchy and the expected hierarchy, this supports evidence that the researcher has a good concept of what is being measured (Boone, 2016) and supports construct validity (Wilson, 2005). Analyzing the relationship between an empirical hierarchy and an expected hierarchy can also function as a tool for questioning assumptions and developing new measurement tools: are the municipalities or municipalities ranked in a way we would expect? Why? Why not? For instance, there are cases in which there is no good match between the empirical and expected hierarchies, which cautions the interpretation of the results and warrants further exploration. Specifically, significant anomalies show scientists when and where to look for a new phenomenon (Kuhn, 1977) and, in turn, present possibilities for discoveries when designing measurements in the social sciences (Fisher & Stenner, 2011).

Furthermore, a good match between the empirical hierarchy and the expected hierarchy not only provides a better understanding of the measurements but also informs further validation steps (Melin & Pendrill, 2023), such as the development of so-called construct specification equations (CSE), which provides a rigorous mathematical and causal conceptualization of the coupling attributes (Pendrill, 2019; Melin, Cano & Pendrill, 2021). This, however, goes beyond the scope of this chapter, but warrants future work to further advance the metrological approach for measurements of social sustainability aspects in municipalities. Additional reading about the role of CSE for validity can be read in the accompanying chapter by Melin (2023).

11.5 A case study: QoL in Swedish municipalities

This section describes the application of the procedure laid out in Sections 11.2–11.4. In the case of social sustainability metrics in municipalities, QoL is a common concept deployed in sustainability policy programs. Even though QoL is a multidimensional construct, our original work included nine dimensions of QoL; our example here is mostly limited to one dimension.

11.5.1 A collaborative process for construct definitions

In 2016, the City of Helsingborg adopted *The Quality-of-Life Program,* which is a municipality-wide policy program aimed at providing residents with QoL. However, after having tried different approaches to evaluate the status and development of QoL, the City of Helsingborg identified the tools they had at hand as insufficient (which initiated the collaboration with RISE). QoL was mainly evaluated by manually analyzing a large set of dimension-relevant indicators, a process that did not yield the information necessary to answer questions identified as relevant by the end-user. The starting point for exploring how the metrological approach could be deployed in developing reasonable and meaningful measurements of social sustainability aspects in municipalities, was thus to obtain an accurate understanding of what questions the end-user – in this case, the City of Helsingborg – wanted to be able to answer. The questions of interest to the municipality were formulated as follows:

i. How is the QoL for residents?
ii. How does QoL differ between people with different background variables such as gender, age, and socioeconomic background?
iii. What prerequisites for QoL does the City of Helsingborg succeed at?
iv. Where is the City of Helsingborg going and how do they know if they are getting there?
v. How does the City of Helsingborg perform compared to other municipalities and cities?

Based on these questions, the objective was to align the construct(s) to be measured according to the user needs. Since it became clear that the focus was not on measuring residents self-perceived QoL, but rather on various factors that constituted preconditions for QoL, the construct of interest was identified as *the municipality's prerequisites for QoL*. The next step was to gain a deeper understanding of the constructs that were to be measured.

In the literature, QoL is often described in broad terms consisting of multiple dimensions. In a policy context, for example, it is conventional to use the WHO's definition of QoL as *an individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards, and concerns* (WHO, 2012: 11). Likewise, many commonly used metrics of QoL have a multidimensional approach, such as the OECD's Better Life Index (OECD, 2022) and Bhutan's Gross National Happiness Index (GNH Centre, 2022), meaning that QoL in a geographical area is estimated by accounting for its performance in several dimensions such as *social relationships* and *personal economy*. This multidimensional view of QoL appeared to align well with the needs identified in dialogue with the City of Helsingborg. As is the case in most municipalities, and indeed most public offices, status reports and evaluations shall serve the function of providing policymakers with a good foundation for making decisions. As such, municipalities are interested in producing and conveying measurement results in a meaningful and understandable manner through established conceptual categories.

Following the multidimensional approach, the intention was to identify the prerequisites essential to a municipality's overall QoL. In order to make the categorization of dimensions as valid as possible, we relied on both previous initiatives and research (OECD, 2022; GNH Centre, 2022) as well as input from workshops and dialogs with the City of Helsingborg. The process resulted in the development of a measurement tool based on nine dimensions explaining a municipality's prerequisites for QoL: *physical health*, *subjective well-being*, *social relationships*, *living situation*, *occupation*, *education*, *environment*, *personal economy*, and *participation*.

After the nine dimensions of prerequisites for QoL were established, a conceptual definition for each category was developed to further specify the relevant features of each dimension. The conceptual definition of *occupation* – the QoL dimension that will serve as the example of when the remaining steps of the procedure is described – was formulated as follows:

> The category occupation concerns what individuals do and have the opportunity to do when it comes to work. Relevant features of occupation are how the work situation looks on an overall level in society, but also how well the work opportunities are distributed between different groups.

As described in Section 11.2, an additional task of interest is to sketch out a construct map describing the characteristics of a municipality being located at different points of the continuum and what it means to go from less to more. At low levels of the continuum, it was suggested that municipalities should be able to provide the most fundamental work-related services, such as keeping their population away from long-term unemployment. Subsequently, as municipalities progress toward the higher end of the same continuum, they should be more successful at more challenging tasks, such as integrating vulnerable groups into the labor market.

As the objective was to measure QoL at the municipality level, an initial database search was conducted to identify potential data sources that provide appropriate indicators. After scanning the identified data sources, we decided to focus on the Swedish database Kolada, a public data hub that gathers regional and municipal indicators from various sources. While the City of Helsingborg had the ambition to perform QoL-related analyses at district levels within the municipality, lack of available and comparable data at district level rendered that approach impossible. Finally, a screening of relevant indicators based on the conceptual definition was performed, resulting in the selection of indicators presented in Table 11.1. To examine invariance – that is; how the indicators function over time – municipality data for all indicators were gathered for 2010, 2015, and 2020.

Table 11.1: Indicator information and categorization of percentages into ordinal scores.

Indicator	Description	0	1	2	3	4
O1	Proportion of refugees (20-64 years) who are working	<40%	40-50%	50-60%	60-70%	>70%
O2	Proportion of citizens (17-24) who are neither working nor studying (reversed)	<85%	85-87.5%	87.5-90%	90-92.5%	>92.5%
O3	Unemployment rate (18-64 years), annual average (reversed)	<90%	90-92.5%	92.5-95%	95-97.5%	>97.5%
O4	Long-term unemployment rate (25-64 years, annual average (reversed)	<92.5%	92.5%-94%	94-95.5%	95.5-97%	>97%
O5	Proportion of foregin-born (20-60 years) who are working	<40%	40-50%	50-60%	60-70%	>70%

Note: Items that were reversed before converted into ordinal scores are marked with (reversed)

11.5.2 Measurement system analysis and measurand restitution

The measurement system was set up as described in Section 11.3, by extending the model ‘man as measurement instrument’ into social sustainability metrics in municipalities. The QoL-related indicators (with difficulties δ_i) constituted the *measurement object*, providing stimulus to municipalities defined as the *instrument* that, in turn, provided a response depending on their ability θ_n, forming the basis for the measurand restitution.

When measuring humans, a person with high ability is more likely to endorse a difficult item than a person with lower ability, and an easy item is more likely to be endorsed by more people than a more difficult item. Similarly, municipalities with high prerequisites for QoL are more likely to succeed in difficult tasks (i.e., score high on difficult items) than municipalities with a lower ability, and easy indicators are more likely to be endorsed by a greater number of municipalities than a more difficult indicator. This is illustrated in Figure 11.2. For instance, a municipality with prerequisites for QoL dimensions Occupation at −2 logits is scored 0 or 1 (Table 11.1) for each indicator, while a municipality with prerequisites for QoL dimension Occupation at 4 logits is scored almost 4 for all indicators.

As outlined in Section 11.3, there is also a need to categorize raw data (i.e., municipality percentages) into a less nuanced polytomous scale. Accordingly, a data categorization table was produced, setting rules for how the raw data percentages were converted into ordinal scores (Table 11.1). The categorization intervals were set up accounting for how well municipalities performed on the different indicators; partly because estimation typically requires a minimum number of 10 observations within each interval, partly because ordinal scores should constitute attainable goals for the municipalities. For example, if the same interval was defined for the work rate among refugees as among the overall population, no municipality would have a realistic chance of reaching even lower ordinal scores.

11.5.3 Theory testing and validation

Data were initially handled in Excel and then transferred into R for categorization of percentages into ordinal scores. Subsequently, the measurement and restitution was conducted using the *RISEkbmRasch* package in R version 4.1.1. (R Core Team, 2019). A summary of the analytic results is presented here:

https://osf.io/g68y7/?view_only=efcf331d40914ae8a0c29294b5f61d08

Initial analyses indicated that the five original items did not seem to measure a single unidimensional construct. First, the eigenvalue in the principal component analysis (PCA) was found to be above the recommended cut-off value of 2.0 (Boone & Staver, 2020). Second, the residual correlation between items O3 and O4 was substantially larger than the relative cut-off value defined as 0.2 over the average correlation (Christensen, Makransky & Horton, 2017). From a theoretical standpoint, the residual correlation is not surprising, as *unemployment rate* and *long-term unemployment rate* are similar concepts that overlap. To deal with the issue of residual correlations, additional analyses were performed, excluding items O3 and O4 individually, to determine how to best proceed. Because the omission of O4 yielded more favorable overall measurement properties (e.g., fewer problems with item fit and residual correlations), it was decided to omit O4 in further analyses.

After omitting item O4, the scale no longer had major problems regarding dimensionality; both eigenvalues in the PCA and the residual correlations were below the recommended cut-off values. However, item O5 (and partly also O1) displayed a slight underfit according to predefined thresholds (Bond & Fox, 2001) but was kept for further analyses because it is a critical part of the construct validity (e.g., integrating vulnerable groups on the labor market) and a reduction to three items would result in a substantial loss of information. Regarding response categories, the category probability curves indicated that indictor categories worked monotonically, consistent with the metric estimate of the underlying construct (Andrich, 1978a; Wright & Masters, 1982). Further examination of *targeting* showed that no major floor or ceiling effects were found, although there were considerable gaps at various points on the scale (Figure 11.2). These gaps could explain why the reliability statistics, as presented in the test information curve, did not reach the recommended levels at any part of the scale. Finally, the invariance of the scale was tested using a DIF analysis to examine whether indicator difficulties varied as a consequence of time and municipality type. Items O1 and O3 showed DIF effects for the time variable above the threshold of 5 logits, whereas O1, O2, and O3 showed DIF effects for municipality type.

As the categorization rule (Table 11.1) is adapted to fit each indicator, the item hierarchy outlined in Figure 11.2 does not provide the objective difficulty of the items. Instead, it displays the difficulty to reach the mean threshold for each item given its own specific categorization rule. This makes it a complex task to pre-specify an item hierarchy because it would require accounting for the specific categorization rule for each indicator. To compensate for this issue, a histogram of the raw scores of each indicator was provided to assess the item hierarchy (i.e., the difficulty of reaching high scores for each indicator). As mentioned in Section 11.5.2, easy indicators are more likely to be fulfilled by a larger number of municipalities than difficult indicators. As such, municipalities' raw scores should be expected to vary depending on the difficulty of the items. For easy indicators, the distribution of municipality raw scores is expected to be close to 100 and vice versa. As illustrated in Figure 11.3, the item hierarchy (from easy to difficult) appears to be O4, O3, O2, O5, and O1. This corresponds well to what was predicted in the construct map, indicating that the most difficult task would be to integrate vulnerable groups into the labor market.

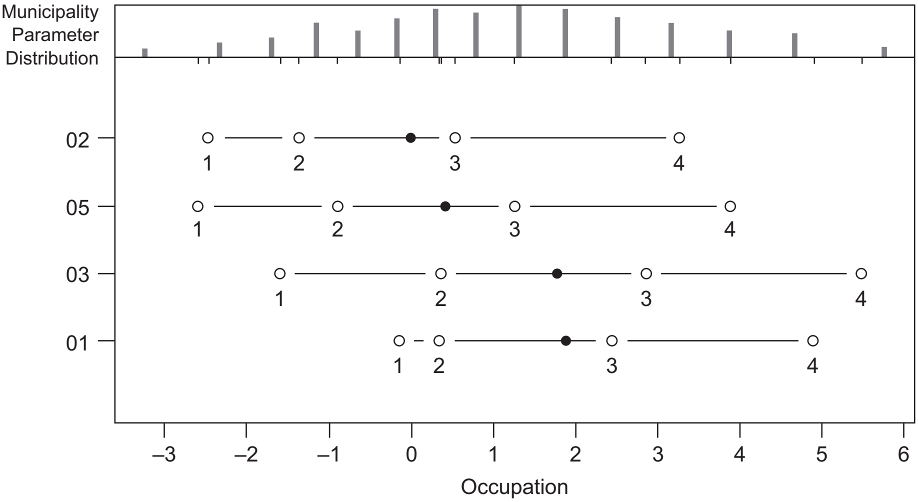


Figure 11.2: Municipality-indicator threshold targeting. At the top histogram, municipalities are located from left to right with lower to higher prerequisites for the QoL dimension Occupation. At the bottom, indicators thresholds are located from left to right with easier to more difficult indicators. White circles indicate the location of thresholds and black circles indicate the mean threshold location for each indicator.

In this section, we have provided a practical example of how the proposed metrological approach for social sustainability metrics can be utilized by developing a measurement for the QoL dimension Occupation. However, as the application of a metrological approach to social sustainability metrics in municipalities is a relatively unexplored area, there are limitations and challenges that must be addressed to fully realize its potential. A problem that became evident was the boundaries imposed by the data currently available at the municipal level. The difficulties of indicators in relation to each other should be invariant, meaning that their difficulty should not be influenced by irrelevant factors such as the time of measurement or the size of the municipality. According to the results in the demonstration, however, this may be a particularly challenging task given the restriction to available data. Indicators currently available at the municipality

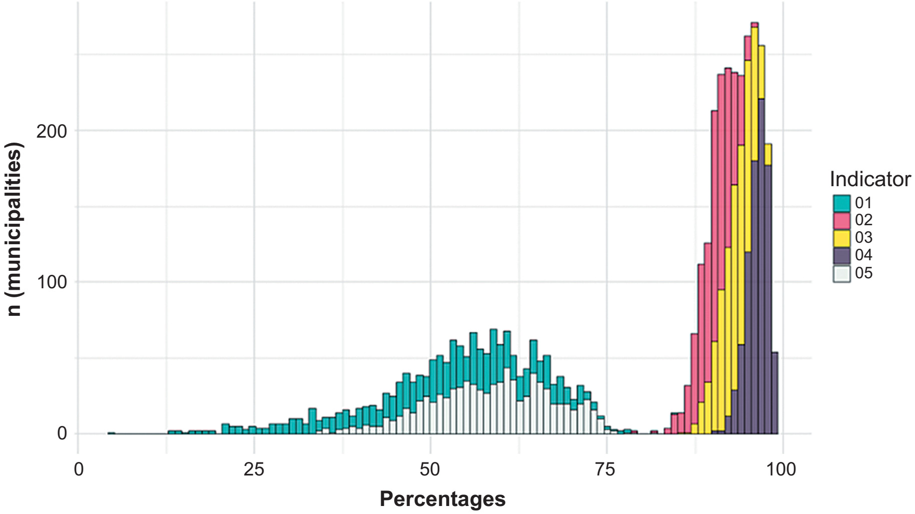


Figure 11.3: A histogram of the raw scores of each indicator used assesses the item hierarchy.

level in Sweden are most likely not developed to produce invariant measurements together with other indicators on the same latent dimension. Another challenge relates to the transformation of raw data (usually presented as percentages) into ordinal scores. If the categorization rule is fully adapted to the performance of the municipalities or other regional entities (e.g., by using percentiles as the basis for the categorization interval), the items will ultimately have the same difficulty level. However, if the same fixed interval is used for all indicators, the intervals may be off-target and result in a loss of information.

11.6 Conclusions

In this chapter, we adopted a metrological approach and presented a process consisting of three main steps (described in Sections 11.2–11.4) for the construction of reasonable and meaningful social sustainability metrics in municipalities. This can be summarized in three bullet points:

i. The first step comprises an iterative process that incorporates both qualitative and quantitative perspectives to define the construct of interest. Because aspects of social sustainability often lack common definitions, it is crucial to develop a proper definition and measurement design for measuring social sustainability. Since measurements are not ends in themselves, it is essential to involve end-users and other relevant experts in the development of useful metrics to ensure that their needs are taken into account. Furthermore, constructing valid and reliable measurements is an iterative process that depends on a carefully thought-out procedure that influences the development of indicators (Boone, 2016). Ideally, indicators should be developed with the intention of capturing different sophistication levels of the underlying construct (Wilson, 2005).

ii. The second step comprises as complete and correct as possible description of the actual measurement system used. With a measurement system where the municipalities act as instruments, the separation between municipality and indicator attributes, as obtained by measurand restitution, allows metrological traceability to be achieved. This ensures both consistency and comparability of social sustainability metrics, which, in turn, allows for meaningful comparisons between different geographical areas and within areas over time. This separation can also serve as a practical tool to inform policies. If municipalities can see themselves on the same scale as the challenges they face (i.e., the indicators making up the scale), they can customize their goals based on their abilities in relation to the difficulty of their challenges.

iii. The final step comprises empirical testing of whether theoretical expectations find support in real-world data. Empirical testing of how the set of indicators works as intended as a unidimensional scale can either validate measurement properties and theoretical expectations or identify significant anomalies that can inform further exploration of the construct of interest.

In the future, it is desirable that sustainability metric systems are coordinated and aligned in networks (Fisher, Melin & Möller, 2021) beyond municipalities. The metrological approach demonstrated here is likely applicable for providing social sustainability metrics at other geographical levels, such as the national, regional, and local levels to enable acting locally while thinking globally. When establishing these networks, it is important that the principles of collaboration, alignment, integration, innovation, and communication guide the process, as in other fields of social sciences (Cano et al., 2019). Specifically, when networks can establish values relative to shared standards, this can open up more fit-for-purpose metrics to be used in developing more meaningful and effective policies and sustainability programs.

Given the high attention paid to social sustainability and the urgent need to evaluate policies and make well-informed decisions to develop new ones, it is time for a "social sustainability metric system" where countries, regions, and municipalities can benchmark where they are relative to where they have been, where they want to go, and what to do next, which would improve pre-conditions for ensuring sustainable societies (Svensson et al., 2022). Thus, building on rigorous conceptual definitions of the attributes to be measured, we have extended the recently developed model 'man as measurement instrument' into social sustainability metrics. Specifically, we encourage a collaborative process between end-users and social sustainability experts and exploit the unique metrological properties of probabilistic conjoint measurement modeling. Demonstrating an approach that uses all three steps together, we have presented principles for models, measurements, and metrology when extending the SI to measurements in the human and social sciences. We believe that this can open up more fit-for-purpose metrics and, in turn, make well-informed decisions to be used in developing more meaningful and effective policies and sustainability programs.

Availability of data

The data that support the findings of this study are openly available in OSF at https://osf.io/g68y7/?view_only=efcf331d40914ae8a0c29294b5f61d08.

References

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Andrich, D. (1978b). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*, 449–460.

Andrich, D. (1988). Rasch models for measurement. SAGE Publications, Inc. https://doi.org/10.4135/9781412985598

Andrich, D. (2013). An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any "threshold disorder controversy. *Educational and Psychological Measurement. SAGE Publications Inc*, *73*(1), 78–124. https://doi.org/10.1177/0013164412450877

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences* (Springer Texts in Education). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-7496-8

Bentley, J. P. (2005). *Principles of measurement systems*. 4th ed., Harlow, England ; New York: Pearson Education.

Berglund, B., Rossi, G. B., Townsend, J. T., & Pendrill, L. (2012). *Measurement with persons: Theory, methods, and implementation areas*. Psychology Press.

Bond, T., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences* (Applying the Rasch Model: Fundamental Measurement in the Human Sciences). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Boone, W. (2016). Rasch analysis for instrument development: Why, when, and how?. *CBE Life Sciences Education*, *15*(4), https://doi.org/10.1187/cbe.16-04-0148

Boone, W., & Staver, J. R. (2020). Principal component analysis of residuals (PCAR). In W. J. Boone & J. R. Staver (Eds.). *Advances in rasch analyses in the human sciences* (pp. 13–24). Cham: Springer International Publishing, https://doi.org/10.1007/978-3-030-43420-5_2

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of pair comparisons. *Biometrika*, *63*, 324–345.

Cano, S. J., Pendrill, L. R., Melin, J., & Fisher, W. P. (2019). Towards consensus measurement standards for patient-centered outcomes. *Measurement*, *141*, 62–69. https://doi.org/10.1016/j.measurement.2019.03.056

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194. SAGE Publications Inc. https://doi.org/10.1177/0146621616677520

E11 Committee. *Guide for measurement systems analysis (MSA)*. ASTM International, https://doi.org/10.1520/E2782-11

Engelhard, G., Jr (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement: Interdisciplinary Research and Perspectives*, *6*(3), 155–189.

Fisher, W. P. (2022). Contrasting roles of measurement knowledge systems in confounding or creating sustainable change. *Acta IMEKO*, *11*(4), 1–6. https://doi.org/10.21014/actaimeko.v11i4.1330

Fisher, W. P., Melin, J., & Möller, C. (2021). *Metrology for climate-neutral cities*. http://urn.kb.se/resolve?urn=urn:nbn:se:ri:diva-57281. (12 September, 2022).

Fisher, W. P. (2020a). Contextualizing sustainable development metric standards: Imagining new entrepreneurial possibilities. *Sustainability. Multidisciplinary Digital Publishing Institute*, *12*(22), 9661. https://doi.org/10.3390/su12229661

Fisher, W. P. (2020b). Measuring genuine progress: An example from the UN Millennium Development Goals. *Journal of Applied Measurement*, *24*.

Fisher, W., & Pendrill, L. (2019). Why metrology? Fair dealing and efficient markets for the United Nations' Sustainable Development Goals. *Journal of Physics*, *7*.

Fisher, W. P., & Stenner, A. J. (2011). Integrating qualitative and quantitative research approaches via the phenomenological method. *International Journal of Multiple Research Approaches*, *5*(1), 89–103. https://doi.org/10.5172/mra.2011.5.1.89

Fors, F. (2012). *Nya mått på välfärd och livskvalitet i samhälet*. Stockholm: Stadsrådsberedningen, Regerionskansliet. https://www.regeringen.se/49b6d0/contentassets/49fa9971dc4d4d6b8dd33c9178a51cc3/nya-matt-pa-valfard-och-livskvalitet-i-samhallet. (11 January, 2023).

GNH Centre. (2022). GNH HAPPINESS INDEX – GNH Centre Bhutan. https://www.gnhcentrebhutan.org/gnh-happiness-index/. (4 November, 2022).

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.). *Measurement and prediction* (Studies in social psychology in World War II. Vol. 4) (pp. 60–90). Wiley.

Hobart, J., & Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: The role of new psychometric methods. *Health Technology Assessment*, *13*(12), https://doi.org/10.3310/hta13120

Johansson, M., Preuter, M., Karlsson, S., Möllerberg, M.-L., Svensson, H., & Melin, J. (2023). *Valid and reliable? basic and expanded recommendations for psychometric reporting and quality assessment*. OSF Preprints, https://doi.org/10.31219/osf.io/3htzc

Kersten, P., White, P., & Tennant, A. (2014). Is the Pain Visual Analogue Scale linear and responsive to change? An exploration using Rasch Analysis. *PLOS ONE. Public Library of Science*, *9*(6), e99485. https://doi.org/10.1371/journal.pone.0099485

Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change*. Revised edition, University of Chicago Press.

Linacre, J. M. (1995). Paired comparisons with ties: Bradley-Terry and Rasch. *Rasch Measurement Transactions*, *9*(2), 425. http://www.rasch.org/rmt/rmt92d.htm

Linacre, J. M. (2000a). Almost the Zermelo model? *Rasch Measurement Transactions*, *14*(2), 754. http://www.rasch.org/rmt/rmt142k.htm

Linacre, J. M. (2000b). Was the Rasch model almost the Peirce model? *Rasch Measurement Transactions*, *14*(3), 756–757. http://www.rasch.org/rmt/rmt143b.htm

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review*, *72*(2), 143–155.

Loftus, P., & Giudice, S. (2014). Relevance of methods and standards for the assessment of measurement system performance in a High-Value Manufacturing Industry. *Metrologia. IOP Publishing*, *51*(4), S219–S227. https://doi.org/10.1088/0026-1394/51/4/S219

Luce, R. D. (1959). *Individual choice behavior*. Wiley.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new kind of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27.

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement*, *51*, 315–327. https://doi.org/10.1016/j.measurement.2014.02.014

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences; Developing a shared concept system for measurement*. 2nd ed., Springer.

Melin, J. (2021). Neurogenerative disease metrology and innovation: The European metrology programme for innovation & research (EMPIR) and the NeuroMET projects. Conference presentation presented at the Pacific Rim Objective Measurement Symposium 2021. https://proms.promsociety.org/2021/.

Melin, J. (2023). Is validity a straighforward concept to be used in measurements in the human and social sciences? *Models, Measurement, and Metrology Extending the SI*.

Melin, J., Cano, S., & Pendrill, L. (2021). The role of entropy in construct specification equations (CSE) to improve the validity of memory tests. *Entropy. Multidisciplinary Digital Publishing Institute*, *23*(2), 212. https://doi.org/10.3390/e23020212

Melin, J., Fisher, W., & Pendrill, L. (2021). A hierarchy of construct theories: Their focus and manifestations. In *Presented at the international objective measurement workshop (IOMW) conference*. Berkeley, CA, https://www.iomw.org/

Melin, J., & Pendrill, L. (2023). The role of construct specification equations and entropy in the measurement of memory. In W. P. Fisher Jr. & S. J. Cano (Eds.). *Person-centered outcome metrology: Principles and applications for high stakes decision making* (Springer Series in Measurement Science and Technology), (pp. 269–309). Cham: Springer International Publishing, https://doi.org/10.1007/978-3-031-07465-3_10

Mohamed, H., & El-Walid, N. (2019). Quality of Life as a Model for achieving sustainable development – an approach study in the light of experiences of some leading countries. *International Journal of Inspiration & Resilience Economy. Scientific & Academic Publishing*, *3*(2), 41–49.

Morel, T., & Cano, S. (2017). Measuring what matters to rare disease patients – Reflections on the work by the IRDiRC taskforce on patient-centered outcome measures. *Orphanet Journal of Rare Diseases*, *12*(1), 171. https://doi.org/10.1186/s13023-017-0718-x

Newby, V. A., Conner, G. R., Grant, C. P., & Bunderson, C. V. (2009). The Rasch model and additive conjoint measurement. *Journal of Applied Measurement*, *10*(4), 348–354.

OECD. (2022). OECD Better Life Index. https://www.oecdbetterlifeindex.org/#/11111111111. (4 November, 2022).

Pendrill, L. (2021). Quantities and units in quality assured measurement. Presented at the Pacific Rim Objective Measurement Symposium 2021. https://proms.promsociety.org/2021/.

Pendrill, L. (2018). Assuring measurement quality in person-centred healthcare. *Measurement Science and Technology*, *29*(3), 034003. https://doi.org/10.1088/1361-6501/aa9cd2

Pendrill, L., & Melin, J. (2019). Measuring counted fractions in healthcare. *TMQ Techniques, Methodologies and Quality*, 60–69.

Pendrill, L. (2014). Man as a measurement instrument. *NCSLI Measure*, *9*(4), 24–35. https://doi.org/10.1080/19315775.2014.11721702

Pendrill, L. (2019). *Quality assured measurement: Unification across social and physical sciences* (Springer Series in Measurement Science and Technology). Springer International Publishing, https://doi.org/10.1007/978-3-030-28695-8

Pendrill, L. (2023). Quantites and units: Order amongst complexity. In *Models, measurement, and metrology extending the SI*. DeGruyter.

Pendrill, L., & Melin, J. (2023). Assuring measurement quality in person-centered care. In W. P. Fisher Jr. & S. J. Cano (Eds.). *Person-centered outcome metrology: Principles and applications for high stakes decision making* (Springer Series in Measurement Science and Technology) (pp. 311–355). Cham: Springer International Publishing, https://doi.org/10.1007/978-3-031-07465-3_11

Pendrill, L. R., & Fisher, W. P. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement*, *71*, 46–55. https://doi.org/10.1016/j.measurement.2015.04.010

R Core Team. (2019). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria, https://www.R-project.org/

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.

Rossi, G. B. (2014). *Measurement and probability [Elektronisk resurs] A probabilistic theory of measurement with applications*. Dordrecht: Springer: Netherlands , http://dx.doi.org/10.1007/978-94-017-8825-0, (11 March, 2022)

Salzberger, T. (2019). *Six decades of measurement using the Rasch model in the social sciences – Setting the agenda for the next sixty years*. Kristianstad.

Simms, L., Zelazny, K., Williams, T., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, *31*(4), 557–566. https://doi.org/10.1037/pas0000648

Svensson, H., Fisher, W. P., Melin, J., Karlsson, S., & Wisén, J. (2022). Initial steps toward measuring and managing prerequisites of quality of life in Sweden. *IMEKO TC* 5.

Thurstone, L. L. (1959). *The measurement of values*. University of Chicago Press.

Toland, M., Li, C., Kodet, J., & Reese, R. (2020). Psychometric properties of the outcome rating scale: An item response theory analysis. *Measurement and Evaluation in Counseling and Development. Routledge*, *0*(0), 1–16. https://doi.org/10.1080/07481756.2020.1745647

Turetsky, V., & Bashkansky, E. (2022). Ordinal response variation of the polytomous Rasch model. *Metron*, https://doi.org/10.1007/s40300-022-00229-w

U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, & U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. (2006). Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims: Draft guidance. *Health and Quality of Life Outcomes*, *4*, 79. https://doi.org/10.1186/1477-7525-4-79

WHO. (2012). The World Health Organization Quality of Life (WHOQOL). https://www.who.int/publications-detail-redirect/WHO-HIS-HSI-Rev.2012.03 (4 November 2022).

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, N.J: Lawrence Erlbaum Associates.

Wilson, M. (2013). Seeking a balance between the statistical and scientific elements in psychometrics. *Psychometrika*, *78*(2), 211–236. https://doi.org/10.1007/s11336-013-9327-3

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*(4), 33–45. https://doi.org/10.1111/j.1745-3992.1997.tb00606.x

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*(12), 857–860.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Trudy Mallinson

12 Extending the justice-oriented, anti-racist framework for validity testing: metrological measurement theory in (re)developing rehabilitation assessments

Abstract: The central thesis of this chapter is that measuring is not a benign act. To unpack this thesis, I consider what it means to measure, starting from the position that measuring in rehabilitation research and practice is more than the application of an assessment tool. Next, I consider how and why measuring in rehabilitation manifests the structural racism that pervades Western and, in particular, U.S. society and is therefore not benign. I consider measuring as an action, something we can mindfully and thoughtfully change in order to reflect more just, inclusive, and diverse perspectives and that can serve to build more just and equitable rehabilitation services. In doing so, I build from the work of Wilson, who describes the construction of measures as occurring in four steps, or as I prefer to consider them, spaces in which mutable measurement actions occur. I consider how the actions that unfold in these concept, item, outcome, and measurement spaces too often marginalize and discount the experiences of minoritized persons. In considering ways measurement actions might be more inclusive, I overview the justice-oriented, anti-racist framework (JAV) approach to building a validity argument proposed by Randall and colleagues in educational assessment research and propose that extending the JAV approach to the probabilistic conjoint measurement RULER reporting framework can support anti-racist measurement in rehabilitation. This extension of the RULER reporting guideline is intended to support transparent reporting of rehabilitation measurement research that includes, elevates, and endorses the experiences of marginalized persons.

Keywords: metrological measurement, justice-oriented, anti-racist, validity, framework rehabilitation assessment

Measuring is not a benign act. Measuring is ***not*** a benign act. Measuring is not a benign ***act***.

The central thesis of this chapter is that measuring is not a benign act. Because measurement ideas and practices are valued – often too uncritically – for their objectivity and simplifications, they can sometimes exert particularly strong – though often

Trudy Mallinson, Department of Clinical Research & Leadership, School of Medicine and Health Sciences, The George Washington University, Washington, DC, USA

https://doi.org/10.1515/9783111036496-012

unintended – effects on the administration of just and equitable policies and programs in rehabilitation, education, and other fields.

That measurement *per se* is not a benign act is not a new idea. In classical measurement engineering – where the Measurement System Analysis approach is also applicable to measurements with persons (section 2.3.3 in Pendrill, 2024) – the "loading effect" is well-known[1] as a modification of the measured object when in contact with the instrument. More fundamentally, in the 1920s, physics was shaken by clear evidence that the process of making observations affects the form of what is observed. This resulted in the need to accept a probabilistic frame of reference for measurement and raised questions that remain unresolved today (Gomez-Marin, 2023). In the study of human behavior, the observer effect, often referred to as the "Hawthorne Effect" occurs when we change our behavior because we know we are being observed (Sedgwick & Greenwood, 2015). Observers may change another's behavior by their presence but may also misperceive others' behavior due to their own expectations or worldview. "Without intending to do so, researchers may encourage certain results, leading to changes in ultimate outcomes" (Street, 2020). Researchers may also make different assumptions about behaviors depending on whether they are reflecting upon their own or others' behavior – referred to as observer-actor bias. We tend to explain others' behaviors as part of their personality, whereas we tend to explain our own behaviors in as arising from outside circumstances and not our own limitations. Thus, there are good reasons to bring a healthy dose of humility to measurement in rehabilitation research beginning with the constructs we choose to measure, the way we design and build measures, and the way we interpret and disseminate results.

To unpack this thesis, that measuring is not a benign act, I will first consider what it means to measure, starting from the position that measuring in rehabilitation research and practice is more than the application of an assessment tool. Next, I consider how and why measuring in rehabilitation manifests the structural racism that pervades Western and, in particular, US society and is therefore not benign. I use the assessment of pain catastrophizing as one example of the problem. Thirdly, I consider measuring as an action, something we can mindfully and thoughtfully change in order to reflect more just, inclusive, and diverse perspectives and that can serve to build more just and equitable rehabilitation services. In considering measurement as an action, I will overview the justice-oriented, anti-racist framework (JAV) approach to building a validity argument proposed by Randall and colleagues in educational assessment research and will propose that extending the JAV approach to the probabilistic conjoint measurement RULER reporting framework can support anti-racist measurement in rehabilitation.

My choice to use pain catastrophizing as an example arose in large part because of my reaction to a recent publication in the *Archives of Physical Medicine and Rehabilitation* that concluded that "Black individuals who have TBI and chronic pain, and

[1] https://electricalguide360.com/loading-effect/

who have public insurance . . . are more likely to cope by catastrophizing, and catastrophizing is related to worse participation outcomes . . . access to care may affect response to chronic pain after TBI" (Sander et al., 2023). Perhaps this uncritical and pejorative conclusion from a nationally respected brain injury research program is unsurprising to Black rehabilitation scholars, and my naïve dismay simply highlights the relative "blindness" I have to such issues because as a white woman, the impact of racism is too often transparent in my everyday experiences. Regardless, I found myself asking why more critical and inclusive psychometric scholarship would not have been undertaken. Concomitantly, a colleague shared an article by Plummer (2021), that exhorted readers to take action, take up space, and have the courage to be heard on issues of racism (Plummer, 2021). And so, I begin this chapter unapologetically, taking the stance that measuring is not a benign act; that measuring can influence the world for good but too often is used to perpetuate oppressive ideologies that not only exclude the perspectives, values, and experiences of those who are not white but to actively categorize others' experiences as less than deviant, and in need of repair without considering that it is the very act of measuring that creates these views; that they are not "revealed" by measuring (Inoue, 2015). My background, as a white female, able-bodied, rehabilitation practitioner, researcher, and mentor, not raised in the USA, is the filter through which I write this chapter and it is to other rehabilitation practitioners, researchers, and early career investigators that I wish to speak to most directly.

It is not lost on me that in describing myself, I mention numerous characteristics that intersect to shape my worldview and the lens through which I write, and yet, in this chapter, I have not directly addressed such intersectionalities. Undoubtedly, Black disabled women are not well served by measures that are ableist and sexist as well as racist. In this chapter, my purpose is to focus on the ways racism infiltrates and impairs rehabilitation measuring.

12.1 What it means to measure

Measuring is far more than the assignment of numbers to categories. There are steps that precede this, which "prepare the ground for measuring" and steps that come later including checking that the assignment of scores was successful and that the measures are used appropriately (Wilson, 2023, p. 7). Assessments are almost always designed with a purpose or use in mind. In rehabilitation, this is usually because we want to distinguish among persons or evaluate change over time in order to make treatment and healthcare decisions. Wilson (2005) notes that this assignment of scores occurs most often "in a practical setting where the results are used to make some sort of decision" (p. 5). Yet the rehabilitation literature is practically silent as to the ways assessments designed to make important healthcare decisions with patients reflect colonial, white, and racist positions or how asking patients to respond to such assessments might enact further microaggressions on persons in a healthcare system that already discriminates and alienates them.

Assessments are designed to capture a single underlying characteristic or trait at a time, though several may be scaled in relation with one another to capture the additional information provided by shared variance (Wilson, 2023). What this trait is, the construct it represents, is an abstraction, made manifest by the creation, that is, the design and development, of the measuring tool. Length is a construct, made manifest by a device such as a ruler or Vernier gauge. Objects have length, but that length is manifest by the structure of the measuring device, by the application of the person using it, and by the interpretation of, or decision made about the obtained length measurement. Still, constructs such as length, time, or mass, have clear internationally recognized definitions and units of measure, international frameworks that ensure measuring devices operate to clear standards, and traceability to the International System of Units (SI). In healthcare assessment, where we attempt to measure human experience, we have yet in most cases to realize such rigor (section 2.4.7 in Pendrill, 2024), and our focus must be on developing conceptually sound, valid, and precise tools if the quality of measurement obtained in health care is to match that obtained in the physical sciences.

Wilson (2023) describes the construction of measures as involving four steps, but I prefer to think of them as spaces. The actions that occur within and between these spaces involve constructions; as a result, our science reflects out those constructions. The constructivist nature of all rehabilitation measurement is important to remember in a time when much is written about the value of team science, the inclusion of diverse community partners' perspectives throughout the design and implementation of research studies, and about reporting out research results in ways community partners understand. When our measures reflect fundamentally racist, colonial assumptions, so does our science and it will undermine all our other attempts at inclusivity. In the concept space, we describe the "thing" that we are trying to measure. Cultural, linguistic, and phenomenological assumptions of the authors, who are almost always white in rehabilitation assessment development, will dominate the construct unless conscious efforts are made to critique those ideas and to actively seek out alternative experiences.

In the item space, text is developed that represents the construct "in the real world" but the question to be asked is "whose real world?" Even among those who speak English (the predominant language in which most rehabilitation assessments are developed), the words chosen and the assumptions expressed within the text of items can be (mis)understood very differently. And for translations to other languages, there is more to be concerned about than the accurate translation of text from one language to another. Are the ideas expressed by the items representative of a broad and diverse range of experiences? While some item content may be explicitly racist, it is important to consider that racist assumptions can be expressed as much by what is excluded from items as what is included.

In the outcome space, the dimension to be addressed is categorized to reflect more or less of the conceptualized/constructed state or trait. Yet, who gets to determine not only what counts as more and less, but also what is valued more or less, is seldom if ever, critically evaluated in rehabilitation research literature. A clear example, from the disability perspective, is the distinction in the FIM rating scale between the highest score of 7, indicating a person can perform an activity "independently" without an assistive device, and a lower score of 6, indicating the person can complete the activity with an assistive device. Disability advocates have argued such distinctions reflect an "ableist" perspective (Bogart & Dunn, 2019). More contemporary perspectives see performance of everyday activities as occurring at the intersection of the person and their environment (Schneidert et al., 2003; WHO, 2001). From this perspective, disability is not a characteristic of the person but describes the intersection between the person and the environment (Mallinson & Hammel, 2010). Thus, a person who can complete an activity, regardless of whether they utilize assistive devices or within an "adapted" environment, is not disabled. This notion is reflected in the more contemporary Section GG functional status items where the highest score of 6 indicates independence, reflecting the person completes the activity, with or without assistive devices. In the same way that rehabilitation assessment, not critically reviewed for ableist perspectives can exclude the experiences of persons with disabilities, so too, rehabilitation assessment, not critically reviewed for racist, "European universal" (Dixon-Román, 2020) assumptions can exclude the experiences of marginalized Black persons.

In the measurement space, data (responses) collected with the items and rating scale steps are evaluated for how well they conform to a probabilistic conjoint measurement model. That is, how the rating scale steps, items, and people cohere to form a unidimensional construct. In this measurement approach, the ordering of items is empirical evidence for the operational definition of the construct. Echoing Gould's (1981), Merry's (2016), Porter's (1995), Powers' (2004), and others' previous efforts, Dixon-Rámon (2020) has made a clear argument for the ways in which racism "haunts" measurement (p. 94). He argues that post-enlightenment colonialism became part of the foundations of science, and that for measurement, "universal Europe and whiteness became that which was scalable and by which values were assigned" (p. 95). Further, he describes how the statistical "tools" we rely on today grew from historical roots that "universalized as European, male, heterosexual, ableist, and Christian, rendering all others as inferior, primitive, or nonhuman" (p. 95). For example, the correlation coefficient developed by Galton has its genesis in the practice of eugenics and Pearson's work aimed to prove the "intellectual superiority of the Aryan race" (p. 95). Statistical methods such as correlation coefficients, principal component analysis, factor analysis, and cluster analysis, all have their roots in comparing humans as same or different and in hierarchically ordered ways, and in doing so, claiming what is preferred (Dixon-Román, 2020; Gould & Rushton, 1997). This is also, and particularly, true of the logistic curve and, by extension, probabilistic models of measurement, which make assumptions about how human differences are categorized and what ordering of those differences is preferred. Critically, how numbers are assigned to reflect "more" or "preferred" reflects underlying social structures, not natural truths.

Having collected data from an assessment and having analyzed the data for alignment to probabilistic conjoint measurement theory, researchers ask the inferential question about the extent to which the ordering of items from "hardest" to "easiest" reflects the proposed underlying construct (Wilson, 2023). It is not always the case that the empirical results match the proposed construct; indeed, many practitioners of probabilistic conjoint measurement consider the ability to identify inconsistencies with the underlying construct and variations in its expression as a strength of the approach (Avlund et al., 1993; Confrey et al., 2021; Sul, 2024; Wilson, 1994). Yet, built within the construct, the items, the rating scale steps, the data collected by the assessment, and the inferences between these, are values about what is socially preferred, and too often this means white. What does it mean if the items do not align with the construct? And who gets to say what it means? What if the items do not align for some people? When and how do we ask if the problem is racist assumptions built into each step of the measurement process?

The propensities of Black Americans to score lower on pain catastrophizing scales and to use "maladaptive" strategies have been repeatedly noted (Meints et al., 2016). But at what point do we ask if the construct itself and the items and rating scales used are biased against non-white respondents and are written to reflect preferred white ways of reacting to pain? When do we challenge the notion that responses to pain categorized as adaptive or maladaptive responses are inherently objective attributes and instead recognize that such categorizations reflect culturally steeped, unexamined, and racist assumptions developed by white researchers within a healthcare system that values white experience?

These questions do not automatically entail a pessimistic perspective on possibilities for improving quantitative communications and applications, but instead point toward the challenging but achievable goals that must be addressed. Contrary to widespread assumptions as to quantitative methods being inherently homogenizing and reductionist (Bryman, 2007; Chwalisz et al., 2008; Merry, 2016; Porter, 1995; Power, 2004), even physical measurements in the natural sciences have been shown to exhibit hierarchically complex irreducible discontinuities across and within communities of research and practice (Blok et al., 2020; Galison & Stump, 1996; Star, 1989; Star & Ruhleder, 1996). The implications of these complexities demand much closer attention to matters of instrument design and construct modeling, but there are rich traditions available to draw from Butz (2018), Fisher (2023), and Wilson (2013).

12.2 How and why measures embody and manifest structural racism

To describe how racist ideas get embedded within rehabilitation assessment and measuring, I begin with a brief history of the construct of pain catastrophizing. I use pain catastrophizing as an example; it is certainly not unique. Rather it serves to illustrate how an idea, developed at another time, uncritically evaluated for its universal European positioning, is used as the basis for developing assessments to measure the construct, which are then tested and validated in ways that reinforce the racist ideas embedded within but not examined, becomes reified in academic literature and is then used in the name of studying health inequities to highlight and point to ways in which Black Americans score poorly on the assessment, and are labelled as more likely to use maladaptive and passive coping strategies. Measuring is the source of the inequities; it does not reveal them.

“Catastrophize” as a term was first used by Albert Ellis in 1962 in a text about “Reason and Emotion in Psychotherapy” (Neblett, 2017) in the context of a male patient being exhorted to “perceive his own tendency to catastrophize” about a perceived sexual failure. It is worth noting that Ellis also admitted to engaging in hundreds of acts of sexual assault on women, sanitized in texts as “non-consensual frotteurism” (Thomason, 2016). A decade later, Spanos and colleagues and Chaves and Brown used the term in the context of susceptibility to hypnosis, describing the concepts of rumination and focused attention on pain (Neblett, 2017). Five years later, Rosensteil and Keefe (1983) developed the Coping Strategies Questionnaire, which consists of three subscales of helplessness, pessimism, and perceived inability to cope with pain (Rosenstiel & Keefe, 1983). A decade later, Sullivan and colleagues published the Pain Catastrophizing Scale with three subscales of helplessness, rumination, and magnification (Sullivan et al., 1995). Around the same time, Vlaeyen and colleagues describe catastrophizing as a patient’s attentional focus on negative aspects of their condition. A decade later, the team reviews literature to describe when catastrophizing occurs, concluding that it occurs when pain is perceived as threatening and the person perceives an inability to cope (Neblett, 2017). In the following decade, based on review of literature, others describe catastrophizing as “misdirected” problem-solving, as avoiding pain-related negative emotions, and goal-directed behaviors that serve a social function. At least through the 1990s, studies of pain catastrophizing seldom, if ever, reported the ethnicity of study participants. Given where these studies occurred (mostly North America), it is likely most study participants were white. The research teams themselves were almost certainly mostly white.

So, when studies examine racial differences in pain catastrophizing, they are fundamentally asking in what ways Black Americans respond to pain in ways that align with the ways white researchers have conceptualized it and the ways white respondents mostly do. It is asking, do Blacks respond to pain in the same ways that are valued by whites? Yet this ignores the omnipresent legacy of slavery, whereby abuse of Black bodies and inflicting atrocious pain on Black persons was routine. It ignores that many older Black Americans still remember relatives who were in slavery, whereas no white person in the USA remembers such a thing. It ignores that even today most medical practitioners believe Black persons experience pain less severely than white persons (Palermo et al., 2023). It ignores that Black Americans experience a litany of microaggressions, not only in daily life, but from heath care systems and practitioners who too often fail to recognize their health needs (Ziadni et al., 2020). In challenging the rehabilitation profession to address racism in rehabilitation practice, Telhan et al. (2020) state that,

> We pride ourselves on treating the whole person and empowering patients to express themselves through their bodies to the best of their ability. But it is long past time to expand that toolbox: to recognize the ways in which the physical body is inextricably tethered to the body politic and to cultivate an awareness of the historical and structural traumas that are mapped onto the lives of so many of our patients. (p. 1842)

As Dixon-Román (2020) notes, psychometric methods were "established based on the falsely assumed privileged access to unobserved mental processes. Not only are the materialized or observed behaviors always-already complicated and differentiated effects of a multiplicity of forces, the interpretations of them cannot be reduced to a universalized epistemology" (Dixon-Román, 2020). An example of this related to pain catastrophizing is the use of prayer. Blacks are much more likely to use prayer as a strategy for coping with pain (Meints et al., 2023). Further, within the Pain Coping Strategies questionnaire, prayer is classified as a "passive" and therefore "maladaptive" or less valuable strategy (Prell et al., 2021). Yet studies demonstrate that prayer can take many forms (e.g., active, neutral, passive) and can be used for many purposes (Upenieks, 2023). From whose perspective is this "maladaptive"? Describing prayer as "passive" is particularly pejorative and negates the experience of many individuals who find prayer not only helpful for managing pain but for dealing with a whole range of life experiences.

Such characterizations seem particularly disrespectful, given that Ellis (who coined the term catastrophizing) claims he cured his fear of women through the "active" solution of approaching women whom he did not know every day in Central Park and making them talk to him (Thomason, 2016). If multiple perspectives are excluded during construct development, when the items and rating scale steps are written, and the validation studies are designed, then analytic methods that were designed to expose and degrade otherness will also replicate the implicit racism and bias within the assessment. Such approaches can only define others as "less than" if their experiences are ignored and the options they are more likely to choose are assigned lower rating scale categories. Pain coping models generally ignore the physiological and psychological effects of racism on marginalized persons, and behaviors frequently used by Black persons to deal with pain may in the context of Black community life, not be maladaptive at all but may in fact be quite effective (Hood et al., 2023). Yet because, from a universalizing European perspective, such behaviors are deemed "maladaptive," they are assigned lower scores, degrading Black behaviors in favor of white (mostly male) responses to pain coping (Booker et al., 2021).

I have described ways in which measuring is not benign; not in the construction of concepts, items, rating scale steps, not in the application of analyses, and not in the interpretation of results that define and delegitimize otherness. Our measuring acts influence the persons who are exposed to the process of measurement, and our measurements can act in ways (however unintended) that commit microaggressions towards persons who already experience too many in their daily lives. I have used pain catastrophizing as an example. But it is just one example.

Articles examining racial differences in pain catastrophizing almost always describe the problem as one of deficit, of "maladaptive strategies" that "blame" the patient, but do not critique the systematic marginalization of Black patients or the systemically universalizing European perspective that defines what is, and is not, "adaptive" (Hood et al., 2023). The field of educational assessment has begun an important discourse on the role of racism in psychometrics, but, with rare exceptions (Balcazar, et al., 3,009), the rehabilitation literature remains acutely silent (Towfighi et al., 2023). To be clear, it is not that the rehabilitation literature does not examine health disparities, it certainly does. For example, in a systematic review, Omar notes that "Black patients are primarily denied access to care, experience lower rates of protocol treatments, poor quality of care, and lack access to rehabilitation" (Omar et al., 2023). The problem, as Omar et al. (2023) clearly identifies, is that studies of racial health disparities in rehabilitation are "disconnected from racism and are displayed as symptoms of a problem that remains unnamed" (p. 1).

I have also argued that we act in each of the four measuring spaces. Through our actions, we can mindfully and thoughtfully change our ways of acting so that our constructions reflect more just, inclusive, and diverse perspectives that hold the potential to support a more equitable healthcare system in the future. Our measures do not reflect immutable truths about the world. We construct the concepts to be measured and the items and rating scale steps and the measurement theory, and the interpretations of results we generate. We construct the data collection methods and the articles we write and publish. At each of these steps, we have opportunities to reflect, to be just, inclusive, anti-racist, and to critically evaluate the impact of our measuring actions. Here, I echo the sentiments of Randall et al. (2002) who note that "ultimately, a commitment to antiracist assessment processes moves beyond simply ensuring representation (of individuals in the field) and includes a fundamental shift in assessment processes" (p. 171).

Randall and colleagues have argued for a justice-oriented, antiracist approach to validation of assessments and measures (JAV). JAV builds on the foundation of critical race theory, which asserts that:

(1) "Race is a social construct, which can be shifted and differentially applied based on the needs of the dominant culture.
(2) Racism is not aberrational; rather it is typical, pervasive, and ingrained in the fabric and system of American society.
(3) It is critical to recognize the relevance of people's everyday lives to scholarship, which includes acknowledging the lived experiences of minoritized peoples and rejecting deficit-informed research that excludes the epistemologies of marginalized groups" (Randall et al., 2022).

These authors highlight our collective responsibility in creating injustices and point to the importance of researcher actions in promoting justice. They point to the work of Young, noting that "Justice is a shared responsibility to which individuals 'contribute by their actions to the processes that produce unjust outcomes' and that our responsibility 'derives from participating in the diverse institutional processes that produce structural injustice' (Young, 2011, p. 105)" (Randall et al., 2022).

So how do we begin to do measuring in rehabilitation without racial bias and in the promotion of justice? Randall et al.'s (2002) answer is that we need to develop "assessment practices specifically designed to sustain, not eradicate, students' cultures, languages, and ways of knowing/being (Inoue, 2015)" (p. 172). Validation should be about demonstrating the value of differences rather than eradicating them. Probabilistic models of measurement, applied uncritically, can harmfully reproduce white hegemony, but applied critically and inclusively, have powerful tools that can be used to highlight and celebrate others' ways of knowing and being (Sul, 2024). Recently colleagues and I published the RULER reporting guideline to promote consistent recommendations for reporting measurement results in a rehabilitation context (Mallinson et al., 2022). We proposed a model with six psychometric domains including conceptual/concept validation, structural validation, external validation, consequential validation, measurement invariance (and reproducibility, reliability), and practical applications and clinical implementation. These guidelines encourage stakeholder engagement throughout the assessment development and refinement process (Mallinson et al., 2022).

In addressing consequential validation, these guidelines referenced "Messick's concern with the uses and consequences of measurement for individual's and society" (Van de Winckel et al., 2022). Although the guidelines were criticized during the development process by reviewers who argued that measurement should be agnostic to such concerns, we the authors nonetheless felt that, in light of arguments from those writing about educational measurement, we clearly failed to address critical issues to ensure rehabilitation measures promote justice and are anti-racist.

The present effort at extending RULER with an adaptation of the JAV framework is accordingly a first attempt to address this shortcoming. Applying a justice-focused anti-racist framework to RULER focuses attention on the ways that uncritical use of probabilistic conjoint measurement theory can "perpetuate injustice and support differential hierarchical power structures (based on race) in society" (Randall et al., 2022). Some rehabilitation researchers may be uncomfortable applying a justice-focused anti-racist approach to measuring. Such a stance requires researchers to reflect on their own practices regarding how they perpetuate whiteness and exclude marginalized others' perspectives from their measuring research (Randall et al., 2022).

In describing what traditional approaches to validity have lacked, Randall et al (2002) provide insights that I believe can inform how a JAV approach can be brought to each of Wilson's four measuring spaces. They note that the essential question is, "Whose LCS patterns are being privileged by an assessment, whose LCS [*linguistic, cultural, substantive*] patterns are being devalued, omitted, suppressed, or marginalized?” (p. 173) or more specifically, “What characteristics of the assessment, the assessment design process, and/or the inferences drawn from the assessment provide evidence of antiracism?” (p. 174).

To encourage rehabilitation researchers to examine, and begin to address, and redress, these questions relative to rehabilitation measurement, I propose that the work of Randall et al, Dixon-Rámon, and others can inform justice-oriented anti-racist considerations in each of Wilson’s four measuring spaces. I then apply the JAV framework, with modest adaptations, to each of the six domains of the RULER framework. As described earlier, Wilson’s insight is that measuring involves acts in all four spaces. Further, he argues for the importance of establishing evidence for the trustworthiness of measuring, which includes precision, reliability, validity (including consequences of measuring), and fairness (Wilson, 2023). But he does not delve into the ways in which injustices, such as racism, can pervade each of the measuring spaces when researchers fail to act in ways that are purposefully just and anti-racist. A reporting guideline, such as RULER (and now RULER + JAV), is in some respects akin to putting the ambulance at the bottom of the cliff because the guideline is applied after an assessment has been designed and tested. While a guideline is not intended to comprise enforceable rules for assessment development and testing, hopefully, knowing research will ultimately be held accountable to such standards prior to publishing will encourage researchers to consider their actions in all four measuring spaces and in their approaches to establishing evidence for trustworthiness.

12.3 Four measuring spaces

12.3.1 Concept space

In probabilistic conjoint measurement theory, all measuring is girded by the notion of a unidimensional trait that persons have more or less of, and which can be observed in the responses to items that represent more or less complexity, challenge, or other dimension of the concept. This is best represented by the construct map. Whether development of the construct map begins with the persons or the items is, according to Wilson, a somewhat arbitrary matter but should be firmly based in deep observational, and I would add phenomenological, research. In an earlier version, he notes that an

> important source of information can be found through . . . participant observation. . . . This can include conversational interviews, recordings of performances, [etc.] . . . used to develop a richer and deeper background for the theory the measurer needs to develop the construct. (Wilson, 2005, p. 53)

Too many rehabilitation outcome assessments lack the kind of deep understandings of the concept being measured that can be obtained on the basis of various methods of participant observation and related phenomenological approaches (Fisher & Stenner, 2011). Still fewer consider the ways in which what Dixon-Rámon (2020) calls "universal European" perspectives are prioritized in developing both the person and item construct maps.

As Wilson (2023) describes, and as I have presented relative to rehabilitation research, measurement concepts are our constructions. What a JAV approach adds is the notion that these are social construct(ion)s, and that we bring our (constructed) identities and those of others to these measuring spaces (Randall et al., 2022). In practice, this means including multiple diverse perspectives in the development of the construct, with clear feedback loops to actively seek out ways in which white perspectives may be dominating what defines "more to less" on the conceptual continuum. It also means being open to the multiple ways of constructing the same concept. If the underlying concept excludes multiple perspectives or does not consider in what ways persons from different racial backgrounds might construct concepts in fundamentally and substantively different ways, then the items, rating scale steps, and analytic approaches will reflect the privileged perspectives of the researchers.

In describing the building of the construct map, Wilson emphasizes the importance of developing the conceptual hierarchy of the people and not just the items. Indeed, a basis of probabilistic conjoint measurement theory is that persons *and* items are hierarchically aligned along the trait. Yet as Randall et al. (2022) note, "These constructs and the reasoning models based on them are necessarily limited by designers' understandings of the network of knowledge, skills, and dispositions that inform these constructs" (p. 173). Thus a JAV approach to the development of the construct map asks researchers to be purposefully mindful in building a diverse team of collaborators and to consider multiple perspectives at the inception of the construct. As Telhan et al. (2020) so clearly state, in reference to racism within rehabilitation research and practice, "The reality is that without an adequate vocabulary to describe the structural trauma facing our patients, we cannot effectively formulate such questions, let alone put our medical expertise and imaginations to work in answering them" (p. 1843). It should no longer be acceptable to comment within the "limitations" section of an article about the lack of diversity in study participants (or study teams for that matter) during development and testing, as if it were an afterthought. Such diversity of perspectives should not be a "plan for future research" – as it too often is but should be a fundamental consideration at the outset of the construction of concepts to be measured.

Consequently, part of the routine steps in constructing the concept should be to ensure persons of color on the research team, in the focus groups, in the conversational interviews, see themselves and their experiences reflected across the entire conceptual continuum. In developing the person construct map, the research team should ask themselves if there is reason to believe that persons of color should not be distributed equally along the continuum. And further, they should ask what changes might be made to the concept such that the perspectives and experiences of persons of color are distributed along the entire continuum. The research team should also be mindful towards genuine and substantive differences in the ways persons experience a given trait that make measuring all persons on the same assessment unfair. That is, measurement researchers and practitioners should remain cognizant that these concepts are construct(ion)s that can be reimagined and revised in service to justice and anti-racism.

As an example, in developing a justice-oriented, anti-racist measure of health literacy, Fleary and colleagues at the Child Health Equity Research Lab began their conceptual development by conducting focus groups with 35 predominantly non-Hispanic Black and Hispanic/Latinx adolescents "to better understand adolescents' definition, operationalization, and use of health literacy" (Fleary & Joseph, 2020). Participants responded to culturally relevant scenarios and responses were coded and used to develop items and response categories for the assessments. Items developed by the research team were "cross-checked with focus groups data for content and consistency with the adolescents' responses and response styles" (p. 4). Next, informed graduate students and practitioners sorted items based on well-established definitions of three types of health literacy. Finally, 17 adolescents, predominantly non-Hispanic Black and Hispanic/Latinx participated in cognitive interviews while they completed the items. "Cognitive interviews results were used to improve (e.g., rewording questions, calibrating the difficulties of the items) or remove problematic items" (p. 4). While this study does not develop the conceptual hierarchy robustly, it does provide a valuable example of how to develop concepts without always starting from white perspectives first and it demonstrates how to effectively involve a variety of community partners throughout the development and testing.

12.3.2 Item and outcome spaces

In moving from construct map to writing items and rating scale steps that will be used as the basis for evaluating persons, measurers must make multiple and frequently arbitrary decisions about matters such as content, wording, format, and respondent burden. Each of these decisions is a conscious act in which the measurer chooses whose LCS is acknowledged, prioritized, and valued. While it is the case, as Wilson (2023) notes, that the items and responses have a "fundamental" relationship to the construct map, it is also the case that "the item is but one of many (often one from an infinite set) that could be used to measure the construct" (p. 73). We almost always have multiple options in choosing not just the item content but the words we use for the items.

For example, the Assessment of Motor and Process Skills (A. G. Fisher, 2003), which evaluates performance on daily living tasks, many of them kitchen tasks, initially included activities such as making a sandwich and only in recent years, added more culturally diverse activities such as making *tostones* or eating an Asian meal with chopsticks. As I noted earlier, the Coping Strategies Questionnaire categorizes prayer as a passive response, and literature suggests Blacks are more likely to identify prayer as a coping strategy for pain, and that "passive" coping strategies are associated with poorer health outcomes. Yet this ignores the scholarship highlighting that prayer, in fact, is practiced in a variety of ways across cultures (Upenieks, 2023). Further, research shows that Blacks do not simply view prayer as a passive strategy (Sharp et al., 2016). Thus, the continued use of an assessment that is outdated and inaccurate in the light of current research is unhelpful and potentially harmful. It may be the case, as for example in the study of coping with pain after brain injury (Sander et al., 2023), that the assessment was included in a national data set at a time when there was less public discussion about racism in assessment and less scholarship about building inclusive, equitable, just assessment. Yet, it is because it is outdated, and because it is included in an ongoing national data set, that the question of how this assessment could be redesigned to be reflect non-white, nondominant perspectives or replaced with a contemporary, anti-racist assessment of pain coping, is so pertinent. Continuing to use such tools in national data monitoring projects perpetuates racial stereotypes and marginalizes persons of color's experiences within rehabilitation research and practice.

Wilson (2023) notes that "the task for the measurer is to choose a finite set of items that represent the construct in some reasonable way" (p. 73) and that "As such, the instrument is a result of a series of decisions that the measurer has made regarding how to represent the construct" (p. 75). As I have argued throughout this chapter, measuring is built on acts that we can choose to reflect predominant "universal European" perspectives and stereotypes of marginalized persons, or not. It is a choice, an act, to continue to use outdated and inaccurate assessments within national data sets, or not. It is a choice, an act, to remain silent on these issues within published manuscripts, or not.

Mislevy (2019) has described how advances in sociocultural psychology are improving educational assessment and measuring. These arguments apply equally to rehabilitation. He describes how LCS patterns impact all kinds of daily activities, including participating in assessments. LCS patterns are "ways of using language and representations, belief systems, and cultural models; and patterns of activity in families, communities, personal interactions, classrooms, and workplaces" (p. 166). He further notes that LCS "patterns and practices," including engaging in assessments, are dependent within history and culture such that performance on (or in response to) assessment items cannot be assumed to only reflect or provide evidence for the underlying trait. This speaks of the need to re-evaluate assessments developed decades earlier to determine if they are still linguistically and culturally relevant and fit for current purposes.

12.3.3 Measurement space

It is worth reminding ourselves of Dixon-Rámon's (2020) work describing how many of the psychometric analytic approaches we rely on, including logits, are based in historical work that was specifically designed to marginalize, degrade, and discount the experiences of persons of color. In large part, reporting of current probabilistic conjoint measurement-based research focuses on describing the hierarchical order of the items, providing detailed fit statistics, principal component analysis of residual, and differential item functioning that point to items that do (or do not) align with the concept as originally constructed. What is too often missing from such reporting, which we tried to address with the RULER guideline, and which Tesio et al. (2023) recently supported, is the inadequate attention paid to the hierarchical ordering of persons. If our conceptual map does not anticipate that persons of color would be more likely to score at the lower end of the scale, as they do for pain catastrophizing or pain coping strategies assessments, for example, we should ask first where the failure(s) in instrument design occurred. We should not assume that some objective and fundamental truth about differences in human performance has been revealed. As Mislevy (2019) notes, "Although psychometrics originated in a quest to measure presumed mental traits in the same sense as length or height, we may instead view these models as tools for managing evidence and inference" (p.173).

An example from the work of Conrad and colleagues (2010) is instructive because it highlights two important issues that can be present simultaneously and which interact to mislead and marginalize (Conrad et al., 2010). The two issues are differential construct definition across groups (difference in the order of item calibrations) and the hierarchical ordering of the persons along the continuum. Conrad et al. (2010) examined a measure of criminality in 7,435 persons screened for substance abuse in the USA. They found important differences in the ordering of items across age and gender, saying:

> The most extreme differences were between adolescent males and adult females. What this means is that adolescent males and adult females had very different hierarchies for crime. More to the point is that, when a sample is predominantly composed of adolescent males as ours was in this study, the measures of adult females will tend to be biased upward. (Conrad et al., 2010)

In this case "upward" means towards greater criminality. If this were also the case, for example, for pain coping strategies, where Black respondents may have different response patterns to whites but are only a small percentage of the overall sample and are more likely to endorse "less adaptive" coping strategies, such as prayer, than whites, then their measures would tend to be biased towards "poorer" coping. Further, when Conrad and colleagues looked at the effect of removing items on which males and females scored differently, they concluded:

> This shows that, indeed, gender does make a difference on certain items in calculating criminality. The problem with dropping items is that the differences in the patterns of crime are real, not

> bias due to bad items. In other words, the items work well within groups, but not when the groups are pooled. (Conrad et al., 2010, p.108)

The point here is that while it is convenient for rehabilitation researchers and practitioners to use the same assessments across all persons and to assume they work in the same ways for everyone, the reality is more complex. Writing about educational outcomes assessment, Fisher, Oon, and Benson (2021) note that "Coherent meaningfulness that does not silence but celebrates the voices of those whose learning outcomes are measured requires close attention to the relational processes by which words and concepts come to represent things in the world" (pp. 5–6). Researchers must engage in an ongoing and reflexive dialogue with the data and with marginalized community partners to determine when observed differences are the result of poorly constructed items, which, by design, excluded the perspectives of those who are not white.

When there are real differences in how groups of individuals experience the world, these must be examined, not erased. In the Conrad et al. study, the concept of criminality, as reflected in the assessment, is predicated on the idea that some crimes are more serious than others, and that all respondents are equally likely to engage in less severe or more severe crimes based on the amount of underlying trait they possess – which is clearly not the case. Conrad et al. demonstrated that females "operationalize" criminality differently than men. It is worth noting that Conrad et al. also examined differences in item hierarchy based on race and did not find any meaningful differences. This may mean that there are indeed, no differences; however, the authors do not describe any work to ensure the construct definition, items, and rating scale steps resonate with the LCS patterns of marginalized persons.

These findings align with Mislevy's (2019) notions of the ways that LCS patterns and practices influence performance on assessments. Not only do women, as a group, not see criminality as men do, individual women bring specific LCS patterns and practices from their homes and communities to bear on assessment responses. The ways in which persons are arrayed in a distribution based on assessment measures, and how they align, or not, with the hierarchy of items, are inextricably linked not to "truth" or some "objective reality" but to the conceptual construction and the constructed items, rating scale steps, the respondents LCS, and the social and historical structures in which assessments are created, designed, and delivered. Inferences made from assessments will be misleading (or worse) if researchers ignore the multiple ways an assessment does not apply to groups of persons or individuals (Mislevy, 2019). Too many rehabilitation articles fail to even present a person map, let alone articulate the ways in which the alignment of persons does or does not support evidence for the underlying concept.

Reporting detailed analyses of the person construct map is an important step in recognizing the central role of LCS patterns and practices in producing person responses. In rehabilitation, study samples are often quite small, which can make demonstrating robust evidence for validity and examining differences in person and item hierarchies based on race and ethnicity challenging. This fact should encourage us to be humble and cautious in determining when an assessment is ready for use in rehabilitation research and practice. It should encourage us to pay more attention to the possibilities that the assessment may misrepresent experiences of marginalized groups.

In particular, we should pause and ask why we, the researchers, believe it is acceptable to write comments in the limitations section of manuscripts to the effect that "future research should examine the validity of the assessment in a more diverse sample" but still conclude that the assessment is fit for use in current research or practice that, inevitably, involves these excluded persons. Would we, the researchers, feel comfortable concluding that the assessment has evidence of validity in white people and can be used to evaluate white people in clinical practice and research? Why does that statement cause discomfort but concluding that an assessment can be used in research and practice when it did not meaningfully include marginalized persons in the design, development, testing, analysis, interpretation, dissemination, and implementation of results is acceptable? As Telhan et al. (2020) ask, "How will we hold ourselves accountable for addressing [inequities in rehabilitation]?" (p. 1843).

I have considered the ways in which measuring (from design to dissemination and implementation) is a series of acts that can marginalize and exclude the perspectives of marginalized persons when it promotes a universal European perspective. I have also considered how such exclusive measuring acts can result in not only inadequate rehabilitation assessments but can inflict microaggressions and real harm on marginalized persons. I have also described ways in which actions taken in each of the measuring spaces can support inclusive, justice-oriented, anti-racist measurement. In this last section, I consider how (Randall et al., 2022) JAV framework can usefully extend the RULER guideline to assist in critically reflecting on ways that rehabilitation assessment can be justice-oriented and anti-racist.

The purpose of reporting guidelines is to "promote transparent and accurate reporting" of health research (*EQUATOR Network*, n.d.). As Altman and Mohr (2014) point out, "Following internationally accepted generic reporting guidelines helps to ensure that published articles contain all the information that readers need to assess a study's relevance, methodology, validity of its findings and its generalizability" (Moher, Altman, et al., 2014, p. 9). The RULER guideline – or any other reporting guideline for that matter, is not designed to proscribe how development of assessments should proceed, or which analyses should be conducted, or how results should be interpreted (Moher et al., 2014). As Tesio et al. (2023) note, the analytic procedures are always under the control of the researcher, an "operator-dependent method" (p. 11).

Randall et al. (2022, p. 173) make an argument for the conditional nature of validity, that "recognizes that the constructs that underpin an assessment design are themselves social constructs" limited by researchers' perspectives and world views, and, drawing on Mislevy's work, by the LCS and life experiences of those being assessed. "What is construct irrelevant for the assessment designer, may very well be construct relevant for the examinee" (Randall et al., 2022, p. 173). To address these concerns,

Table 12.1: RULER framework of the six psychometric domains.

	Conceptual/content validity	Structural validity	External validity**	Consequential validity	Reproducibility and reliability	Practical applications and clinical implementation
Scope of psychometric domain	Considers the extent to which items represent the underlying construct and the existing evidence that those items reflect the underlying construct. Considers the extent to which the construct is underrepresented by the content of the items or to which assessment responses reflect processes other than the underlying construct. Considers to what extent stakeholders believe the items are relevant to and representative of the domain.	Refers to the extent to which items, rating scale steps/thresholds, and people cohere to form a measure that substantively reflects the assumptions of the Rasch model. The key assumptions of the Rasch model are: – Unidimensionality: items all measure the same underlying trait. – Hierarchical order: harder items are harder for everyone and more able people score higher on all items.	Refers to the extent that measures from the assessment of interest align with measures from similar and/or different assessments. Indicates that performance on the assessment relates to skill/knowledge in the target domain. The extent to which the measures are associated with external factors that are consistent with expectations.	Refers to the extent to which assessment results are suited to the purpose for which they were intended and to the clinical and social impact of using measures from an assessment for clinical and research decision-making.	Refers to the extent to which assessment results are comparable across individuals, time, raters, or settings.	Refers to the extent to which the assessment is feasible, practical, and usable in the settings and for the purposes it is intended to be used

Measurement question addressed	Is the model underlying the construct sound? Is the underlying construct sufficiently represented by item content for the intended purpose of the assessment? Do stakeholders believe the items cover the concept of interest?	What are the internal structure of the items, the rating scale steps, and the persons?	Does the assessment align with similar (and different) assessments in ways that are predicted by theory and/or clinical experience?	What is the intended purpose of the assessment and what kind of decisions are the measures intended to inform?	Does the assessment produce measures that are reproducible and consistent?	Is the assessment feasible, practical, and usable in the settings in which it is intended to be used? Is the assessment fit for the purposes for which it is intended?
Evidence	Theoretical or logical or empirical rationale for the trait. As a result of Rasch Analysis, items and rating scale steps are ordered from "lowest" to "highest" in a way that aligns with the theoretical rationale or clinical understanding.	Rating scale step structure: Step thresholds ordering is theoretically and clinically reasonable for interpretation and intended use. Unidimensionality: Items are sufficiently unidimensional such that any dimensionality does not disrupt person measures.	Item calibrations from current assessment align with other assessments sharing similar underlying trait.	The measures obtained from the assessment are useful for helping clinicians, patients, and caregivers (and other stakeholders) make good treatment/clinical decisions that inform rehabilitation services, research, and policy.	Reproducibility: Step thresholds and item calibrations are reproducible across samples/groups/settings, time, raters, and forms.	The assessment can be used as intended (not shortened or altered) in the rehabilitation setting for which it was intended (e.g., clinical practice, research, policy). Training to administer assessment is available, affordable, and practicable.

(continued)

Table 12.1 (continued)

	Conceptual/content validity	Structural validity	External validity**	Consequential validity	Reproducibility and reliability	Practical applications and clinical implementation
	Empirical evidence that stakeholders believe the items are relevant and representative of the domain, including expert opinion, Delphi testing, or cognitive testing	Measurement accuracy: Score properties and interpretations generalize to and across individuals, groups, settings, and tasks. Iterations of analysis: Successive iterations of Rasch analyses improve assessment psychometrics.	Person measures demonstrate an association with measures (or raw scores) from assessments sharing similar (or different) underlying trait.	The measures do not result in unintended consequences that negatively impact the lives of persons with disability including but not limited to reducing timely access to appropriate rehabilitation services.	Reliability: The degree of association among item scores and/or total raw scores obtained at different times, by different raters, or different assessment forms.	The assessment fits in clinical workflow. The assessment is in languages that are appropriate for the intended uses.

Justice-focused anti-racist Validity Proposition	Marginalized stakeholders are engaged in initial construct definition, and consideration of causality and inferences between construct and item responses. For existing assessments, marginalized stakeholder perspectives are sought on construct and alignment of items and steps. Construct mapping specifically addresses non-white experiences and perspectives.	Items and rating scale steps are designed and evaluated on the broadest range of persons, specifically seeking to identify and redress ways in which white perspectives may dominate. Analysis purposely seeks to identify ways marginalized stakeholder linguistic, cultural, and substantive patterns can be reflected in item content and rating scale options. Considers how racist ideas are codified within the hierarchy of rating scale steps, items, and people.	Conceptual and structural validity of the alternative criteria/ assessment is critiqued and evaluated for inclusivity of perspectives and values. Expectations regarding association of assessment measures with external criteria are developed in collaboration with non-white, marginalized stakeholders.	Results are interrogated for ways in which interpretations of assessment results reinforce racial stereotypes. Interpretation of results specifically examines the clinical and social impact of using measures for clinical and research decision-making on marginalized stakeholders.	Assumptions about sameness and difference are clearly explicated. Alternative experiences and ways of knowing are acknowledged. Ways in which marginalized stakeholders are expected to respond in alignment with the European universal are questioned and critiqued.	Consideration of ways that measurement is not a benign act and might be inflicting further microaggressions within heath care environments that have historically treated and currently treat black and brown persons inadequately. Evaluation of disrupting current practices for marginalized stakeholders is addressed. For example, consideration of who is administering the assessment and how are addressed.

(continued)

Table 12.1 (continued)

	Conceptual/content validity	Structural validity	External validity**	Consequential validity	Reproducibility and reliability	Practical applications and clinical implementation
Questions addressing JAV proposition*	Are marginalized stakeholders involved at every stage of the construct definition and refinement? Is the construct being measured specifically designed to be inclusive? Whose values, perspectives, ways of knowing, and experiences does the construct reflect, normalize, or marginalize? Is the construct explicitly anti-racist? Are the specific false and oppressive narratives it seeks to disrupt specifically stated?	Do the items reflect/ reify negative stereotypes of minoritized populations? Are there items that actively disrupt negative stereotypes about minoritized populations? Has antiracist content been explicitly integrated into items? Does the content/ language of the items privilege a particular linguistic or cultural way of thinking/ making sense of the world?	How are criterion variables selected? Does this selection process consider the history/ impact/legacy of white supremacist hegemonic practices? Does this process of criterion selection seek to disrupt these hegemonic practices? Have alternative criterion variables – those that center and reflect the values of non-white respondents been considered/ examined?	Do test/assessment results serve to further marginalize already minoritized populations? How does/can structural racism impact the results of this assessment? What groups may be advantaged and disadvantaged and/or privileged by the administration of this assessment? In the short term? The long term?	Are Eurocentric ways of knowing and processing information being privileged over other ways of knowing and processing information? Have we considered/ interrogated whether or not the assessment requires responses that are in alignment with the dominant white discourse?	What historical logics of testing and racism are respondents bringing to the rehabilitation assessment situation? For observational assessments, have rater severity/leniency and bias been examined, with particular consideration of race of observer and respondent?

		How have values shown up in the items/tasks? And which social identity groups do these values reflect? How stable is the construct across social, cultural, and racial contexts?	To what extent do test-criterion relationships generalize across historically marginalized populations?	What systems would have to be in place for respondents to be successful on this assessment? Are current systems rooted in white supremacist values?	Have a wide range of interpretations been considered that acknowledge the different ways of knowing, thinking, and experiencing of Black respondents? Do the items allow for multiple ways of thinking, knowing, and doing?	Do researchers make their own racial perspectives visible, rather than just describing races of respondents?
Evidence	Theoretical or logical or empirical rationale for the concept/trait that specifically challenges the European universal. Considerations of causality and inference are determined by inclusion of a broad range of marginalized stakeholders.	Hierarchical ordering of persons is specifically investigated for racial bias. Lack of inclusion of sufficient marginalized stakeholders is stated as a major study limitation. When racial differences in responses are found, assessment tool is critiqued, not racial group.	Criterion assessments are critiqued for racial biases. Evidence regarding generalizability across marginalized populations is sought and critiqued. Failure to evaluate generalizability across marginalized stakeholders is stated as a major study limitation.	When racial differences on assessments are found, discussion of results considers how assessment marginalizes minority respondents rather attributing to personal/group deficits. Discussion considers how assessment can be improved for marginalized patients.	Considers whose values and what inequities/inequalities are being reproduced and/or concealed. Discussion acknowledges how reliability methods seek alignment to social constructions and in doing so, may perpetuate racial inequities.	Considerations about social response biases are described. For observational assessments, racial differences between assessors and respondents are clearly described. Researcher's racial background is explicitly stated.

*Questions addressing JAV propositions: adapted from Randall (2022).

**External validity: adapted from Randall (2002) and Dixon-Román (2020).

†Reliability and reproducibility: adapted from Dixon-Román (2020).

Randall et al. present a table of six elements of validity criteria and related questions that attempt to expose and overcome racist perspectives embedded within traditional approaches to validity.

In Table 12.1, I adapt the work of Randall and colleagues, which was developed in the field of educational assessment, to reflect rehabilitation outcome measurement concerns. There is much overlap, as the fundamental concerns within both fields have a good deal in common; both are, after all, measuring people. The six criteria presented by Randall and colleagues align to a large extent with those of the RULER organizing framework presented by Mallinson et al. (2022). Both construct articulation and content criteria from JAV align with aspects of conceptual/construct validity of RULER. JAV internal structure aligns with RULER structural validity and JAV relations to other variables aligns with RULER external validity. JAV response process aligns to some extent with RULER reproducibility and reliability but also with some aspects of RULER structural validity. JAV does not directly consider implementation in the same way that RULER does.

In Table 12.1, I extend the six psychometric domains of RULER with three considerations: a justice-focused anti-racist validity proposition, questions addressing the JAV proposition, and evidence for the validity of the JAV proposition related to the specific psychometric domain. For the justice-focused anti-racist validity propositions, I have tried to reflect the ways in which Randall and colleagues, Dixon-Rámon, Mislevy, and others have described what inclusive anti-racist measurement looks like both in terms of developing new assessments and in using and revalidating existing assessments. For the questions addressing each proposition, I have tried to stay true to the words and intent of Randall and colleagues, modifying the content where necessary to reflect the concerns of probabilistic conjoint measurement theory and/or rehabilitation research.

Because the JAV and RULER domains do not completely align, I chose to locate JAV questions within the RULER domain they seemed, to me, to best address. Others may disagree with these choices and further dialog is welcomed. For evidence, I attempted to provide examples of the output, results, and interpretations that could be reported as demonstrating a justice-oriented anti-racist position. These are just a few examples; undoubtedly better ones exist. If they encourage authors to reflect deeply on how structural racism is persisted through assessment, they have served their purpose. This extension of the RULER reporting guideline is intended to support transparent reporting of rehabilitation measurement research that includes, elevates, and endorses the experiences of marginalized persons. Work by Fisher, Oon, and Benson (2021) presents ways that metrological measurement extends to different levels of complexity. In the case of rehabilitation, we often focus on the level of the individual patient. But the principles of measurement described here can extend to service lines, healthcare organizations, or geographical regions and the measurement and justice-oriented considerations apply across each of these levels of complexity.

There are claims that the SI could become the universal language for all sciences if extensions were made to the concept of measurement (Jeckelmann & Edelmaier, 2023). It has also been argued by proponents of probabilistic models of measurement that such models produce psychometric results that are equivalent to tools such as rulers, scales, thermometers, governed by rules to assure precise calibration of such tools. It is indeed idealistic to wish that measures of human experience be treated with as much care as rulers and scales. In this chapter, I have argued that the concerns of healthcare measurement broadly, and rehabilitation measurement specifically, are more foundational, because the construction of measures requires us to consider the essential humanity of those who will be measured by our tools.

In response to Telhan et al.'s (2020) question, one step we can take in holding ourselves accountable for addressing inequities in rehabilitation research and practice is to assume that measuring is not a benign act, and that we can and should do better in how we design, use, and interpret assessments.

References

Avlund, K., Kreiner, S., & Schultz-Larsen, K. (1993). Construct validation and the Rasch model: Functional ability of healthy elderly people. *Scandinavian Journal of Social Medicine*, *21*(4), 233–246.

Balcazar, F. E., Suarez-Balcazar, Y., Taylor-Ritzler, T., & Keys, C. B. (2009). *Race, culture and disability: Rehabilitation science and practice*. Jones & Bartlett Publishers.

Blok, A., Farias, I., & Roberts, C. (Eds.). (2020). *The Routledge companion to actor-network theory*. Routledge.

Bogart, K. R., & Dunn, D. S. (2019). Ableism special issue introduction. *Journal of Social Issues*, *75*(3), 650–664. https://doi.org/10.1111/josi.12354

Booker, S. Q., Bartley, E. J., Powell-Roach, K., Palit, S., Morais, C., Thompson, O. J., Cruz-Almeida, Y., & Fillingim, R. B. (2021). The imperative for racial equality in pain science: A way forward. *The Journal of Pain*, *22*(12), 1578–1585. https://doi.org/10.1016/j.jpain.2021.06.008

Briggs, D., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, *4*(1), 87–100.

Bryman, A. (2007). Barriers to integrating quantitative and qualitative research. *Journal of Mixed Methods Research*, *1*(1), 8–22.

Butz, M. R. (2018). *Chaos and complexity: Implications for psychological theory and practice*. CRC Press.

Chwalisz, K., Shah, S. R., & Hand, K. M. (2008). Facilitating rigorous qualitative research in rehabilitation. *Rehabilitation Psychology*, *53*(3), 387–399.

Confrey, J., Shah, M., & Toutkoushian, E. (2021). Validation of a learning trajectory-based diagnostic mathematics assessment system as a trading zone. *Frontiers in Education: Assessment, Testing and Applied Measurement*, *6*(654353). doi: 10.3389/feduc.2021.654353

Conrad, K. J., Riley, B. B., Conrad, K. M., Chan, Y.-F., & Dennis, M. L. (2010). Validation of the Crime and Violence Scale (CVS) against the Rasch measurement model including differences by gender, race, and age. *Evaluation Review*, *34*(2), 83–115. https://doi.org/10.1177/0193841X10362162

Dixon-Román, E. (2020). A haunting logic of psychometrics: Toward the speculative and indeterminacy of blackness in measurement. *Educational Measurement, Issues and Practice*, *39*(3), 94–96. https://doi.org/10.1111/emip.12375

EQUATOR Network. (n.d.). Retrieved November 19, 2023, from https://www.equator-network.org/about-us/

Fisher, A. G. (2003). *Assessment of motor and process skills: Volume 1: Development, standardization, and administration manual*. (5th ed.), Three Star Press.

Fisher, W. P., Jr. (2023). Measurement systems, brilliant results, and brilliant processes in healthcare: Untapped potentials of person-centered outcome metrology for cultivating trust. In W. P. Fisher Jr. & S. Cano (Eds.), *Person-centered outcome metrology: Principles and applications for high stakes decision making* (pp. 357–396). Springer. https://link.springer.com/book/10.1007/978-3-031-07465-3

Fisher, W. P., Jr, Oon, E. P.-T., & Benson, S. (2021). Rethinking the role of educational assessment in classroom communities: How can design thinking address the problems of coherence and complexity? *Educational Design Research*, *5*(1), 1–33

Fisher, W. P., Jr, & Stenner, A. J. (2011). Integrating qualitative and quantitative research approaches via the phenomenological method. *International Journal of Multiple Research Approaches*, *5*(1), 89–103.

Fleary, S. A., & Joseph, P. (2020). Adolescents' health literacy and decision-making: A qualitative study. *American Journal of Health Behavior*, *44*(4), 392–408. https://doi.org/10.5993/AJHB.44.4.3

Galison, P., & Stump, D. J. (1996). *The disunity of science: Boundaries, contexts, and power*. Stanford University Press.

Gomez-Marin, A. (2023). David Bohm's unfinished revolution. *Science*, *381*(6657), 489–489. https://doi.org/10.1126/science.adi3423

Gould, S. J. (1981). *The mismeasure of man*. W. W. Norton (Reprinted 1996 in revised and expanded edition).

Gould, S. J., & Rushton, J. P. (1997). The mismeasure of man: Revised and expanded edition. *Society*, *34*(3), 78–82.

Hood, A. M., Morais, C. A., Fields, L. N., Merriwether, E. N., Brooks, A. K., Clark, J. F., McGill, L. S., Janevic, M. R., Letzen, J. E., & Campbell, L. C. (2023). Racism exposure and trauma accumulation perpetuate pain inequities-advocating for change (RESTORATIVE): A conceptual model. *The American Psychologist*, *78*(2), 143–159. https://doi.org/10.1037/amp0001042

Inoue, A. B. (2015). *Antiracist writing assessment ecologies: Teaching and assessing writing for a socially just future*. The WAC Clearinghouse; Parlor Press. https://doi.org/10.37514/PER-B.2015.0698

Jeckelmann, B., & Edelmaier, R. (Eds.). (2023). *Metrological infrastructure*. De Gruyter Oldenbourg. https://doi.org/10.1515/9783110715835

Kaiser, F. G., & Wilson, M. (2000). Assessing people's general ecological behavior: A cross-cultural measure. *Journal of Applied Social Psychology*, *30*(5), 952–978.

Mallinson, T., & Hammel, J. (2010). Measurement of participation: Intersecting person, task, and environment. *Archives of Physical Medicine & Rehabilitation*, *91*(9 Suppl), S29–33. https://doi.org/S0003-9993(10)00284-4 [pii] 10.1016/j.apmr.2010.04.027

Mallinson, T., Kozlowski, A. J., Johnston, M. V., Weaver, J., Terhorst, L., Grampurohit, N., Juengst, S., Ehrlich-Jones, L., Heinemann, A. W., Melvin, J., Sood, P., & Van de Winckel, A. (2022). Rasch reporting guideline for rehabilitation research (RULER): The RULER statement. *Archives of Physical Medicine and Rehabilitation*, *103*(7), 1477–1486. https://doi.org/10.1016/j.apmr.2022.03.013

Meints, S. M., Illueca, M., Miller, M. M., Osaji, D., & Doolittle, B. (2023). The pain and PRAYER scale (PPRAYERS): Development and validation of a scale to measure pain-related prayer. *Pain Medicine (Malden, Mass.)*, *24*(7), 862–871. https://doi.org/10.1093/pm/pnad020

Meints, S. M., Miller, M. M., & Hirsh, A. T. (2016). Differences in pain coping between black and white Americans: A meta-analysis. *The Journal of Pain*, *17*(6), 642–653. https://doi.org/10.1016/j.jpain.2015.12.017

Merry, S. E. (2016). *The seductions of quantification: Measuring human rights, gender violence, and sex trafficking*. University of Chicago Press.

Mislevy, R. J. (2019). Advances in measurement and cognition. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 164–182. https://doi.org/10.1177/0002716219843816

Moher, D., Altman, D. G., Schulz, K. F., Simera, I. & Wager, E. (Eds.). (2014). *Guidelines for reporting health research: A user's manual*. Wiley. https://www.wiley.com/en-gb/Guidelines+for+Reporting+Health+Research%3A+A+User%27s+Manual-p-9780470670446

Neblett, R. (2017). Pain catastrophizing: An historical perspective. *Journal of Applied Biobehavioral Research*, *22*(1), e12086. https://doi.org/10.1111/jabr.12086

Omar, S., Nixon, S., & Colantonio, A. (2023). Integrated care pathways for black persons with traumatic brain injury: A critical transdisciplinary scoping review of the clinical care journey. *Trauma, Violence & Abuse*, *24*(3), 1254–1281. https://doi.org/10.1177/15248380211062221

Palermo, T. M., Davis, K. D., Bouhassira, D., Hurley, R. W., Katz, J. D., Keefe, F. J., Schatman, M., Turk, D. C., & Yarnitsky, D. (2023). Promoting inclusion, diversity, and equity in pain science. *Canadian Journal of Pain*, *7*(1), 2161272. https://doi.org/10.1080/24740527.2022.2161272

Pendrill, L. R. (2024). Quantities and units: Order amongst complexity. In *Models, measurement, and metrology extending the SI – Trust and quality assured knowledge infrastructures: Vol. Chapter 2*. De Gruyter.

Plummer, D. L. (2021, October 19). Leading in the post-Floyd era. *Age of Awareness*. https://medium.com/age-of-awareness/leading-in-the-post-Floyd-era-322d73469c12

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Power, M. (2004). Counting, control, and calculation: *Reflections on Measuring and management.Human Relations*, *57*(6), 765–783. DOI: 10.1177/0018726704044955

Prell, T., Liebermann, J. D., Mendorf, S., Lehmann, T., & Zipprich, H. M. (2021). Pain coping strategies and their association with quality of life in people with Parkinson's disease: A cross-sectional study. *PLoS ONE*, *16*(11), e0257966. https://doi.org/10.1371/journal.pone.0257966

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting white supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment*, *27*(2), 170–178. https://doi.org/10.1080/10627197.2022.2042682

Rosenstiel, A. K., & Keefe, F. J. (1983). The use of coping strategies in chronic low back pain patients: Relationship to patient characteristics and current adjustment. *Pain*, *17*(1), 33–44. https://doi.org/10.1016/0304-3959(83)90125-2

Sander, A. M., Christensen, K., Loyo, K., Williams, M., Leon-Novelo, L., Ngan, E., Agtarap, S., Martin, A. M., Neumann, D., Hammond, F. M., Hanks, R., & Hoffman, J. (2023). Coping with chronic pain after traumatic brain injury: Role of race/ethnicity and effect on participation outcomes in a TBI model systems sample. *Archives of Physical Medicine and Rehabilitation*. https://doi.org/10.1016/j.apmr.2023.03.003

Schneidert, M., Hurst, R., Miller, J., & Ustun, B. (2003). The role of environment in the international classification of functioning, disability and health (ICF). *Disabil Rehabil*, *25*(11–12), 588–595. https://doi.org/10.1080/0963828031000137090QM6N3XQF2DYKL7CN [pii]

Sedgwick, P., & Greenwood, N. (2015). Understanding the Hawthorne effect. *BMJ*, *351*, h4672. https://doi.org/10.1136/bmj.h4672

Sharp, S., Carr, D., & Panger, K. (2016). Gender, race, and the use of prayer to manage anger. *Sociological Spectrum*, *36*(5), 271–285. https://doi.org/10.1080/02732173.2016.1198948

Star, S. L. (1989). The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving. In M. Huhns & L. Gasser (Eds.), *Distributed artificial intelligence* (pp. 37–54). Morgan Kaufmann. (Reprinted in Bowker, G., Timmermans, S., Clarke, A. E., & Balka, E. (Eds). (2015). *Boundary objects and beyond: Working with Leigh Star*. MIT Press, pp. 243–259.)

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, 7(1), 111–134. (Reprinted in Bowker, G., Timmermans, S., Clarke, A. E., & Balka, E. (Eds). (2015). *Boundary objects and beyond: Working with Leigh Star*. MIT Press, pp. 377–415.)

Street, F. (2020, August 17). *The observer effect: Seeing is changing*. Farnam Street. https://fs.blog/observer-effect/

Sul, D. (2024). Situating culturally specific assessment development within the disjuncture-response dialectic. In W. P. Fisher Jr. & L. Pendrill (Eds.), *Models, measurement, and metrology extending the SI* (pp. in press). De Gruyter.

Sullivan, M. J. L., Bishop, S. R., & Pivik, J. (1995). The pain catastrophizing scale: Development and validation. *Psychological Assessment*, *7*(4), 524–532. https://doi.org/10.1037/1040-3590.7.4.524

Telhan, R., McNeil Ba, K. M., Lipscomb-Hudson, A. R., Guobadia, E. L., & Landry, M. D. (2020). Reckoning with racial trauma in rehabilitation medicine. *Archives of Physical Medicine and Rehabilitation*, *101*(10), 1842–1844. https://doi.org/10.1016/j.apmr.2020.07.001

Tesio, L., Caronni, A., Simone, A., Kumbhare, D., & Scarano, S. (2023). Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. *Disability and Rehabilitation*, 1–14. https://doi.org/10.1080/09638288.2023.2169772

Thomason, T. C. (2016). The shadow side of the great psychotherapists. *Counseling & Wellness: A Professional Counseling Journal*, 5. https://openknowledge.nau.edu/id/eprint/2346/

Towfighi, A., Boden-Albala, B., Cruz-Flores, S., El Husseini, N., Odonkor, C. A., Ovbiagele, B., Sacco, R. L., Skolarus, L. E., & Thrift, A. G., & American Heart Association Stroke Council; Council on Cardiovascular and Stroke Nursing; Council on Cardiovascular Radiology and Intervention; Council on Clinical Cardiology; Council on Hypertension; Council on the Kidney in Cardiovascular Disease; and Council on Peripheral Vascular Disease. (2023). Strategies to reduce racial and ethnic inequities in stroke preparedness, care, recovery, and risk factor control: A scientific statement from the American Heart Association. *Stroke*, *54*(7), e371–e388. https://doi.org/10.1161/STR.0000000000000437

Upenieks, L. (2023). Unpacking the relationship between prayer and anxiety: A consideration of prayer types and expectations in the United States. *Journal of Religion and Health*, *62*(3), 1810–1831. https://doi.org/10.1007/s10943-022-01708-0

Van de Winckel, A., Kozlowski, A. J., Johnston, M. V., Weaver, J., Grampurohit, N., Terhorst, L., Juengst, S., Ehrlich-Jones, L., Heinemann, A. W., Melvin, J., Sood, P., & Mallinson, T. (2022). Reporting guideline for RULER: Rasch reporting guideline for rehabilitation research: Explanation and elaboration. *Archives of Physical Medicine and Rehabilitation*, *103*(7), 1487–1498. https://doi.org/10.1016/j.apmr.2022.03.019

WHO. (2001). The international classification of functioning, *Disability and Health – ICF*.

Williams, J. (2010). Audit and evaluation of pedagogy: Towards a cultural-historical perspective. In T. Rowland & K. Ruthven (Eds.), *Mathematical knowledge in teaching* (pp. 161–178). Springer.

Wilson, M. R. (1994). Comparing attitude across different cultures: Two quantitative approaches to construct validity. In M. Wilson (Ed.), *Objective measurement: Theory into practice, Volume 2* (pp. 271–294). Ablex.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Wilson, M. R. (2013). Seeking a balance between the statistical and scientific elements in psychometrics. *Psychometrika*, *78*(2), 211–236.

Wilson, M. (2023). *Constructing measures: An item response modeling approach* (2nd ed.), Routledge. https://doi.org/10.4324/9781003286929

Young, I. M. (2011). Responsibility for Justice. Oxford: Oxford University Press.

Ziadni, M. S., Sturgeon, J. A., Bissell, D., Guck, A., Martin, K. J., Scott, W., & Trost, Z. (2020). Injustice appraisal, but not pain catastrophizing, mediates the relationship between perceived ethnic discrimination and depression and disability in low back pain. *The Journal of Pain*, *21*(5–6), 582–592. https://doi.org/10.1016/j.jpain.2019.09.007

Linda Morell, Sean Tan, and Mark Wilson

13 Aligning and disentangling science content and practices: the relationship between measures of twenty-first-century skills and the content underlying them

Abstract: *A Framework of K-12 Science Education* (Framework; National Research Council, 2012) represented a significant shift from previous conceptions of science learning. The Framework provided the principles for a new conceptual framework for science education and a set of science standards. Researchers and practitioners have grappled to understand the connections among the elements of the framework. These elements are called science *dimensions* and include (a) disciplinary core ideas (DCIs), (b) scientific and engineering practices (SEPs), and (c) crosscutting concepts (CCCs), which are more fully described in the publication called the *Next Generation Science Standards* (NGSS; NGSS Lead States, 2013). The most significant new element in the Framework is the *performance expectation*, which describes how a DCI, an SEP, and a CCC must be combined to guide teachers about what students should know and be able to do along each of the dimensions. This study, using the Berkeley Evaluation & Assessment Research Assessment System methodology, extends and builds on studies about learning progressions in scientific practices (specifically, scientific argumentation) and science content (understanding of interdependent relationships in ecosystems). Analyzing data from 1,387 middle and high school students in a large and diverse urban school district in the United States, we find evidence that the content and practice can indeed be aligned and disentangled with the use of theoretically robust learning progressions. Conducting a thorough investigation of learning progressions of this nature aligns with, rather than contradicts, the feasibility of adopting a shared conceptual framework for measurement across disciplines, as proposed by Mari et al. (2023).

Keywords: assessment, science assessment, educational assessment, learning progressions

Acknowledgments: The authors are grateful for the efforts of many who made this work a reality, including Weeraphat Suksiri, PhD, for help with the assessment design, data collection, and preliminary analysis while he was a graduate student at UC Berkeley; Jonathan Osborne at Stanford University, Anna MacPherson at the American Museum of Natural History, and Sara Dozier at the California State University (Long Beach) for their contribution to design and development of material; and Karen Tran of the SERP Institute for coordinating the project. Finally, the authors are very grateful to the teachers and their students who provided data for the study.

Linda Morell, Sean Tan, Mark Wilson, Berkeley School of Education, University of California, Berkeley

https://doi.org/10.1515/9783111036496-013

13.1 Introduction

In education in general, and in cognitive studies in particular, the articulation of how to relate success in argumentation with command of the underlying knowledge, which is the basis for the arguments involved has been a long-standing conundrum. Indeed, in say, science education, for many years the recommended curriculum has been structured into (a) science content and (b) science inquiry. In particular, this has been manifested in the development of science assessments that separately measure (a) a student's command of science concepts and (b) the student's ability to use scientific practices such as argumentation to reason about those science concepts (e.g., see NAEP, 2019). Yet, real science does not separate content from practice; in fact the two are always being used together – how to achieve this in a measurement context has been challenging. In this chapter, we examine the practical resolution of this conundrum using a principled approach to measurement that allows one to align and disentangle measurements across content and argumentation, using the science domain as the context. Further, we leverage the affordances of scientific measurement principles to align the skill and the context using a common metric, which can be seen as a micro-example in support of the conceptual framework for measurement as discussed by Mari et al. (2023).

Assessing a skill or competency typically involves using stimulus material situated within a given context. Therefore, it is crucial to understand the connection between the skill and the context or content domain. In this chapter we provide an example of a twenty-first-century skill, argumentation, situate it within the science context of ecology for middle school students, and investigate the relationship between the two using the Berkeley Evaluation & Assessment Research (BEAR) Assessment System or BAS in short. Our investigation employs learning progressions and multidimensional Rasch modeling to delve into and disentangle the connection between the skill and content.

Ever since *A Framework of K-12 Science Education* (referred to as the *Framework* in this chapter henceforth; National Research Council, 2012) provided the principles for a new conceptual framework for science education and a set of science standards, researchers and practitioners have grappled with the connections between the parts of the framework. These elements are called science *dimensions*[1] and include (a) disciplinary core ideas (DCIs), (b) scientific and engineering practices (SEPs), and (c) crosscutting concepts (CCCs), which are more fully described in the publication called the

[1] Note that *dimension* will be used in two different ways in this chapter. We use it here in this paragraph in terms of the three NGSS science "dimensions"—the DCIs, SEPs, and CCCs. Later, we also use the term "dimension" as it is used in psychometrics, where it is used to describe the unobserved student variables in the data generating model—often also called "constructs" in psychometrics. We make this distinction because use of the term amongst people from different disciplines has caused confusion.

Next Generation Science Standards (NGSS; NGSS Lead States, 2013). These dimensions are an important organizing tool to understand the Framework. However, the most significant new element in the Framework is the *performance expectation* (PE), which describes how a triple of a DCI, an SEP, and a CCC must be combined to guide teachers about what students should know and be able to do along each of the dimensions. These PEs are:

> Grounded in situated cognition theory (Greeno et al., 1996), which contends that knowledge is situated, being part of a product of activities, and context and culture-dependent, based on the environment in which it is developed and used (Brown et al., 1989). Situated cognition emphasizes real-world situations that are meaningful for learners to make sense of phenomena and motivates them to figure out solutions for problems in new contexts actively. (He et al., 2023)

Two examples from chemical reactions are illustrated in Table 13.1:

> MS-PS1–2 is a PE of students' knowledge-in-use statement using the SEP of analyzing and interpreting data and the CCC of patterns, whereas MS-PS1–5 is another knowledge-in-use PE using the SEP of developing and using models and the CCC of energy and matter. (He et al., 2023)

There are four DCIs, seven SEPs, and eight CCCs, and together they are combined as a selected subset of all possible triples to form some 208 PEs (NGSS Lead States, 2013). This represents a very dense and complexly structured conceptual mapping of the topic-contents of science through K-12 education. As such it is, on the one hand, a boon to anyone wanting to know the details of what students should be learning when following the NGSS approach to science education. On the other hand, it is a formidable challenge for any science teacher trying to keep track of and comprehend the progress of their students as they learn.

This then is the **first** challenge that we address in this chapter – how to organize these many PEs into a set of structures that can be usefully deconstructed back to the PEs, but that nevertheless embody a developmental framework that helps a teacher assess where their students are in their expected progress, and to plan for their next instructional steps to move the students along the path to success. The approach we describe here uses the concept of a learning progression (LP). LPs provide a way to conceptualize learning as being on a continuum and have been described by Duschl and colleagues (2011) as "the successively more sophisticated ways of thinking about a topic that can follow one another as children learn about and investigate a topic over a broad span of time." Now more than a decade later, learning progressions endure and flourish as a tool to guide standards and assessment (Smith et al., 2006), formative assessment practices (Alonzo & Elby, 2019), and support the generation of curriculum Wiser and Smith, 2008; Smith et al., 2004). Thus, this first challenge has already been addressed in the literature, and, as we shall see below, researchers from the BEAR Center have played an important role in accomplishing that.

Moreover, as noted above, the PEs that are the foundation of these LPs are deeply integrated across critically different types of constructs, the DCIs, SEPs, and CCCs –

Table 13.1: Two performance expectations (PEs) from the NGSS framework (adapted from table 1 in He et al., 2023).

Task name	Learning performance	NGSS performance expectation
T1. Gas-filled balloons	Students analyze and interpret data to determine whether substances are the same based upon characteristic properties.	MS-PS1-2. Analyze and interpret data on the properties of substances before and after the substances interact to determine if a chemical reaction has occurred.
T2. Layers in a test tube	Students construct a scientific explanation about whether a chemical reaction has occurred by using patterns in data on the properties of substances before and after the substances interact.	MS-PS1-5. Develop and use a model to describe how the total number of atoms does not change in a chemical reaction and thus mass is conserved.
T3. Battery under water	Students develop a model of a chemical reaction that explains the regrouping of atoms form new substances and that mass is conserved.	

hence the **second** challenge is to design the LP-based assessments in such a way that, even though the observations of the science performances that will be observed are indeed integrated across the dimensions, they can be used to create separate indicators of each of the different components – that is, into the relevant DCI, SEP, and CCC. We will illustrate this challenge in this initial paper using just two of the three dimensions, a DCI and an SEP.

In this chapter, we explore two learning progressions (Wilson, 2009, 2023) – one in the NGSS DCI of life science (interdependent relationships in ecosystems) and the other in the science and engineering practice of engaging in argument from evidence (scientific argumentation) – to understand their own structure and the relationship between them in order to inform the ongoing conversation about learning progressions in science assessment. Specifically, we are interested in investigating the question, *what is the relationship between the content* (DCI) and *practice* (SEP) constructs*?* Building on previous empirically validated learning progressions (Dozier et al., 2023; Osborne et al., 2016), it seems clear that an investigation into their relationship is an important contribution to research on learning progressions.

This chapter is organized into four sections. In the first section, we describe previous research upon which this chapter is based. Next, we describe the methodology and processes used to empirically investigate the DCI and the SEP. Third, we provide the empirical findings from our investigation. Specifically, we test the dimensionality (unidimensionality versus multidimensionality) of the constructs. Finally, we review the findings and discuss their implications.

13.2 Previous research

Among the myriad publications on learning progressions in science (Kaldaras et al., 2023; Jin et al., 2019; Jin et al., 2017; Hokayem & Gotwals, 2016; Gotwals, 2012; Songer et al., 2009), we have identified three key studies below that are especially informative of the strategy and tactics we use in our paper: Dozier et al. (2023), Osborne et al. (2016), and Yao et al. (2015).

13.2.1 Validating a learning progression for student understanding of ecosystems

The first study that we focus on proposed and examined evidence for the validity of a learning progression for student understanding of the interdependent relationships in ecosystems (Dozier et al., 2023). The study builds on previous work by using innovative assessments that are designed as context-rich tasks situated around natural phenomena, that use a variety of engaging selected-response questions or item types, and that target conceptual understanding as specified in the NGSS (NGSS Lead States, 2013). The data collected from 1,366 middle school students in a large and diverse urban school district in the United States was used to investigate the validity of the learning progression.

The learning progression in the study comprised four waypoints (Wilson, 2023). The least sophisticated waypoint (Waypoint 0) is named "Notions" and indicates that students could only express naive and often inaccurate knowledge about ecosystems. The next waypoint (Waypoint 1) indicates that students understand direct relationships in nature. For example, students at this waypoint understand the predator-prey relationship and are able to predict the effect of a change in the size of one population on the size of another population. Next, students at Waypoint 2 understand indirect relationships in ecosystems and can predict population changes among organisms that are one or more steps removed from each other in a food web for example. Finally, students at Waypoint 3 understand complex relationships among organisms in an ecosystem such that they can predict changes in more than two components in an ecosystem based on microscopic or macroscopic populations or the availability of resources.

The study describes the use of Wilson's (2023) construct modeling approach called the BAS. Using this approach, researchers developed assessment material including questions or items within tasks or item bundles (Rosenbaum, 1988) and collected empirical evidence to validate the learning progression. Each task contained multiple questions, and each question was designed to map to one specified waypoint of the learning progression.

The authors were careful to analyze and present data from only the content (ecosystems) learning progression, although they did mention that questions requiring scientific argumentation thinking were also included in the tasks. Many of the items included in the validation of the ecosystems learning progression study as well as the argumentation items mentioned by the authors are used in this chapter.

13.2.2 Validated learning progression for argumentation in science

In the second study, Osborne and colleagues (2016) focused their work around investigating how middle grade students argue from evidence. The authors posit that "Argumentation is a central feature of science" (Osborne et al., 2016, p. 821), and is of the utmost importance to study. This is reflected in the prominence given to it in the Framework and NGSS.

The authors introduced a hypothesized three-tiered learning progression for scientific argumentation, which is based on the Toulmin model (Toulmin, 1958). Toulmin's (1958) model begins with (a) a claim as a conclusion whose merit we seek to establish, which is supported by (b) the evidence in the form of data that supports the claim, and (c) a warrant, which is the link between the claim and evidence that forms the foundation of justification for the initial claim. According to Osborne et al. (2016), Toulmin's practical model "has formed the basis of many schemas used in research analyzing student discourse (Cavagnetto, 2010; Erduran et al., 2004; Zohar & Nemet, 2002) because of its relative simplicity."

Osborne and colleagues based their hypothesized learning progression on Toulmin's model, extended it into the science space, and identified four waypoints in the progression in sophistication of the argumentation. Specifically, they situated scientific argumentation within the physical science domain of the structure of matter. At the lowest waypoint, a student demonstrates no facility with argumentation. At the next waypoint, the student can construct their own claim and also identify another person's claim. It is at this waypoint that the student can also support a claim with a piece of evidence and identify another person's piece of evidence. At the next higher waypoint, the student can construct a warrant, identify a warrant, construct a complete argumentation, and provide an alternative counterargument. At the highest waypoint, the student can provide a counter-critique, construct a one-sided comparative argument, and construct a counter claim with justification. The authors conceptualize scientific argumentation as a competency that obviously "demands a complex orchestration of construction and critique of claims, warrants, and evidence in situations that require scientific knowledge to solve" (Osborne et al., 2016, p. 826). Findings from the study suggest that the hierarchical structure of the hypothesized scientific argumentation learning progression is supported by the empirical evidence collected.

According to Osborne and colleagues, scientific argumentation is not simply an aptitude "that can be assessed, but rather, a competency which draws on a mix of content knowledge, procedural knowledge, and epistemic knowledge" (Osborne et al., 2016, p. 823) as discussed in the *OECD Science, Technology and Industry Outlook* (OECD, 2012). The authors' intent when designing items was to emphasize students' ability to engage in reasoning in a scientific context. For this purpose, they provided the relevant content knowledge which students could use to engage in argumentation in order to minimize the prerequisite domain-specific knowledge.

Argumentation can be both a novel and a demanding cognitive activity for students depending on the complexity of the argument. The authors acknowledged that scientific argumentation relies on content knowledge, which has implications for developing assessment tasks to tap both the content and the practice.

13.2.3 Investigation of science content and practice

In the third study, Yao and colleagues (2015) laid out a methodology to empirically investigate middle school students' science content knowledge and competency in a scientific practice. The paper investigated the function of the items designed to assess science content knowledge in physical science (states of matter) and items designed to assess the science practice of argumentation. The researchers describe their investigation, which included the use of a multidimensional framework. Dimensionality analyses were performed to investigate whether the relationship between the science content and practice conformed to the anticipated test design.

The science test examined by the authors contained items mapped to two different learning progressions. A set of content-related tasks and items that were designed to assess middle school students' knowledge of changes of state (e.g., solids, liquids, and gas) appeared on the test. This content is commonly taught in eighth grade science and is a component of a larger learning progression in physical science called the Structure of Matter. The test also included a set of tasks and items designed to assess argumentation contextualized within the changes of state domain.

The study found that the multidimensional between-item[2] model fitted the data better than either the unidimensional model or the multidimensional within-item[3] model. This result validated the test design because the content and argumentation items were designed to address different constructs and the scoring schemes for the study were designed to foster an emphasis on either the content or the argumentation learning progression.

[2] In between-item models, each item is an indicator of only one underlying construct.

[3] In within-item models, items may be indicators of more than one underlying construct, in this case *both* the content and the argumentation constructs.

13.2.4 Appreciation of prior research

Each of the key studies produced empirical evidence relevant to our current purpose. The first study provided a validated learning progression of how students understand the interdependent relationship in ecosystems, but it did not include an analysis of scientific argumentation situated within the ecosystem context. The second study produced a learning progression for scientific argumentation, but it was situated in a physical science context and may not necessarily apply to other contexts (in particular, for the life science context we are exploring). Similarly, the third study investigated the relationship between the physical science content domain of changes of state and the practice of scientific argumentation, but the findings may not extend to other science domains (e.g., relationship between a life science content domain and scientific practice). That said, each study contributed to our thinking and prompted us to undertake this study because we could leverage the learning progressions from the first two studies and the methodology from the third study.

13.3 Methods

A total of 1,387 middle and high school students in a large and diverse urban school district in the United States responded to assessment items delivered through an online platform called the BEAR Assessment System Software (BASS; Wilson et al., 2019). All students provided active assent to participate in the study in accordance with the university's Committee for the Protection of Human Subjects (CPHS protocol 2010-09-2,241). Demographic information for the school district indicates that the student body consisted of Latino (33%), Asian (30%), White (14%), African American (7%), and other [American Indian (<1%), Filipino (4%), Pacific Islander (<1%), multiracial (7%), declined to state (4%)] identifying students. Approximately 27% of the students are designated as English language learners, 13% are in special education, and 52% are categorized as socioeconomically disadvantaged.

The construct modeling approach developed by Wilson (2023) guided the development of the study. Given the robustness of this approach, it has been used to investigate the validity of a variety of science assessments (e.g., Dozier et al., 2023; Chi et al., 2022; Morell et al., 2017), and was recommended in *Developing Assessments for the Next Generation Science Standards* (National Research Council, 2014). Specifically, we used the BAS (Wilson, 2023) to empirically investigate the nature of the relationship between the science content and practice. The BAS is a comprehensive assessment framework used to develop and validate learning progressions and assessment material. It includes four building blocks (see Figure 13.1): the construct map, items design, outcome space, and Wright map. Each building block is used in the development cycle to ensure high quality assessments.

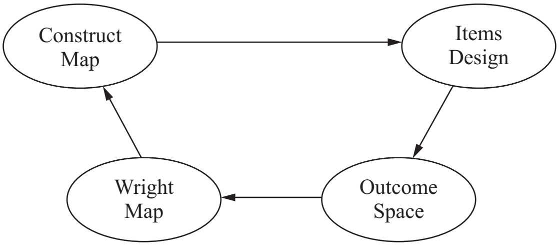


Figure 13.1: The BEAR Assessment System (Wilson, 2023).

For this study, we were able to leverage the learning progressions, items within tasks, and scoring schemes that were validated previously (Dozier et al., 2023; Osborne et al., 2016), so that we could focus on modeling the relationship between the learning progressions.

13.3.1 Construct maps

In this study, we use the term "construct map" synonymously with "learning progression" but understand that there can be multiple relationships between the two (see, for example, figure 8, Wilson, 2009, p. 725). The construct map's most important features are that there is a coherent and substantive definition for the construct's content and that the construct is composed of an underlying developmental continuum that can be expressed by waypoints of increasing sophistication described in terms of (a) respondent characteristics and (b) characteristics of item responses (Wilson, 2023). For this study two construct maps were used. Figure 13.2 shows the construct map for student understanding of the independent relationships in ecosystems, and Figure 13.3 shows an updated construct map of student competency in scientific argumentation.

As can be seen, the construct map in Figure 13.2 describes student understanding from least sophisticated (Waypoint 0, Notions) at the bottom to most sophisticated (Waypoint 3, Complex relationships) at the top. The construct map was based on previous research (Hayes et al., 2017; Gotwals & Songer, 2013; Chandler, 1992; Hokayem & Gotwals, 2016; Gotwals & Songer, 2010; American Association for the Advancement of Science, 2001) and designed by Dozier and colleagues (2023).

The construct map in Figure 13.3 developed and investigated by Osborne and colleagues (2016) was based on Toulmin's argumentation framework. The image in Figure 13.3 differs from the previously published image in that the waypoints go from least sophisticated to most sophisticated, here. This was done in keeping with Wilson's vision of displaying the construct map.

Given that the two learning progressions were already developed and validated, we used them to develop and identify items.

Waypoint	Description
	Complex Relationships
3	Students predict changes in more than two components in an ecosystem based on changes in microscopic populations or available resources.
	Indirect Relationships
2	Students predict the effects of change in one population on another population with an indirect relationship.
	Students predict the effects of availability of and competition for resources (e.g., food, space, water, shelter, and light) on populations.
	Direct Relationships
1	Students predict the effect of a change in the size of one population on the size of another population in mutual, commensal, or parasitic relationships.
	Students predict the effect of a change in the size of one population on the size of another population in a predator-prey relationship.
	Students predict the effects of change in plant populations throughout the food web using the knowledge that plants form the base of the food web and are living organisms.
	Notions
0	Students express naïve knowledge about ecosystems.

Figure 13.2: Learning progression of understanding interdependent relationships in ecosystems (Dozier et al., 2023).

13.3.2 Items design and outcome space

Following the BAS, after the construct map is initially defined, it is operationalized through the next two building blocks. The items design and outcome spaces are the second and third building blocks in the BAS. The items design is the systematic design of questions intended to elicit a response from the responder that can be mapped back to the waypoints of the construct map. Each item is designed to tap into the student's knowledge, skill, and attributes so that it can be located on the construct at a particular waypoint. The specific item responses are mapped back to the construct map via the outcome space.

For the science assessment administered to students, a total of 51 items were used – 31 items designed to assess student understanding of interdependent relationships in ecosystems, and 20 items assessing student competency in scientific argumentation. All items were selected response-type items presented in a variety of formats including multiple choice, drag and drop, fill in the blank, matrix, and sorting. Items appeared in six different item bundles or tasks named Succession, Foxes, Lion,

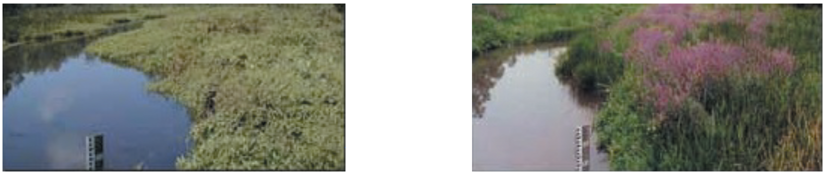
Waypoint	Description
3	This waypoint marks the top anchor of the progress map. The student explicitly compares and contrasts two competing arguments, and also constructs a new argument in which they can explicitly justify why it is superior to each of the previous arguments.
	Student makes an evaluative judgment about two competing arguments and makes an explicit argument (claim + justification) for why one argument is stronger and why the other is weaker (claim + justification).
	Student critiques another's argument. Fully explicates the claim that the argument is *flawed* and justification for why that argument is flawed.
2	Student offers a counter argument as a way of rebutting another person's claim.
	Student makes a claim, selects evidence that supports that claim, and constructs a synthesis between the claim and the warrant.
	Student identifies the warrant provided by another person.
	Student constructs an explicit warrant that links their claim to evidence.
1	Student identifies another person's evidence.
	Student supports a claim with a piece of evidence.
	Student another person's claim.
	Student states a relevant claim.
0	No evidence of facility with argumentation.

Figure 13.3: Learning progression for scientific argumentation (adapted from Osborne et al., 2016).

Whales, and Invasive. Each task contained between 8 and 13 items mapping to either the ecosystems or the argumentation construct map.

Figure 13.4 shows the prompt for a task named "Invasive" and two questions – one that maps to Waypoint 1 of ecosystems (content) construct map and one that maps to the argumentation construct map (Waypoint 1). The task is named "invasive" because it shows a before and after picture of a location – before the purple loosestrife (an invasive species) is introduced, and after the purple loosestrife is introduced into the ecosystem. The invasive task contains five questions designed to tap a student's competency in scientific argumentation and three questions designed to provide evidence for one waypoint in the content (ecosystems) construct map. Note that each question within the larger task was designed to map to just one waypoint of one construct map. Although science dimensions (e.g., DCIs, SEPs, and CCCs) are routinely mixed together in curriculum and instruction, recall that, following the arguments presented above, for assessment of the respective learning progressions, it is essential to keep them separate. This means that one can provide targeted feedback to teachers and students about what a student knows or can do within a given science dimension unclouded by other science dimensions.

The purple loosestrife, a wetland plant, was imported to North America from Europe. The purple loosestrife has spread to many wetland ecosystems in the United States.



Before the introduction of purple loosestrife

After the introduction of purple loosestrife

Observations by scientists:

- Observation #1: The purple loosestrife is a plant that grows twice as fast as the winged loosestrife (a native plant in these wetlands)
- Observation #2: The winged loosestrife is a plant that has 10 different species of insect that eats its leaves. The purple loosestrife has 3 different species of insects and 2 species of birds that eat its leaves.
- Observation #3: Snakes have been observed underneath both the winged loosestrife and purple loosestrife. The snakes eat bird eggs.

Example of an Ecosystems Waypoint 1 Item

Which is an example of an herbivore in this ecosystem? (N1)

A. Winged loosestrife
B. Purple loosestrife
C. Insect
D. Snake

Example of an Argumentation Waypoint 1 Item

Sophie says: I *think the purple loosestrife is a successful invader because there are very few herbivores that eat it.*
What evidence listed below supports **Sophie's** claim? (N10)

A. There are only 5 species that eat the purple loosestrife compared to 10 species that eat the winged loosestrife.
B. The purple loosestrife grows twice as fast as the native winged loosestrife.
C. Snakes eat the eggs of birds under the loosestrife.

Figure 13.4: Invasive task's prompt and two items – one connected to the ecosystems construct map and the other connected to the scientific argumentation construct map.

13.3.3 Wright maps

The fourth building block of the BAS is the Wright map. This building block summarizes how inferences about student understanding are made from the student responses. Specifically, this step is where the values from the outcome space are translated back to match the framework of the construct map, and hence from numbers back to interpretation.

To understand the data and answer the research question, "*What is the relationship between the content (interdependent relationships in ecosystems, ECO) and practice (scientific argumentation, ARG) constructs?*", we followed the procedure laid out by Yao and colleagues (2015). The multidimensional random coefficient multinomial logit (MRCML) framework (Adams et al., 1997; Wang et al., 1997) was used. The MRCML framework is "flexible and permits the estimation of various Rasch-type models" (Yao et al., 2015, p. 6). The analysis examines validity evidence in this case focusing specifically on internal structure (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014).

Choosing a model to analyze the psychometric properties of the assessment is important to understanding the internal structure of the test contents and respondents. Rasch models (Rasch, 1960/1980) provide both rich and convenient ways to model item difficulties and person proficiencies on the same scale.

Figures 13.5–13.7 show graphical representations of the models used to explore the relationship between the ecosystems and scientific argumentation constructs.

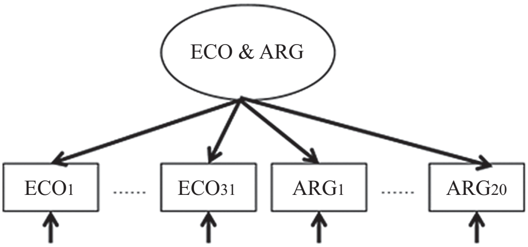


Figure 13.5: Graphical representations of the first model, the unidimensional model.

As shown in Figure 13.5, under the unidimensional model, it is assumed that the content (ECO) and the practice (ARG) measure a single ability. In the unidimensional model, the items are regarded as assessing content knowledge and facility with the practice together.

Underlying the use of the multidimensional between-item model (as shown in Figure 13.6), the content and practice are assumed to measure separate (content and practice) constructs, respectively. For the study by Yao and colleagues (2015), the multidimensional model represents the rationale of the test design so "that the argumentation items were designed to minimize the demand for the domain-specific knowledge and emphasize students' ability to reason" (Osborne et al., 2016). This holds true for the items used in this study also. An advantage of using the multidimensional model instead of two unidimensional Rasch models (one for each construct) is that parameters can be calibrated for the whole science test simultaneously. As mentioned by Yao and colleagues (2015), when "the underlying dimensions are not orthogonal . . . [as was assumed in their case and ours] students' performance on a subset of items measuring one dimension provides collateral information about their performance on a different subset of items measuring another dimension" (p. 17). As noted by Wang and colleagues (1997), the estimation accuracy of the multidimensional between-item model is gener-

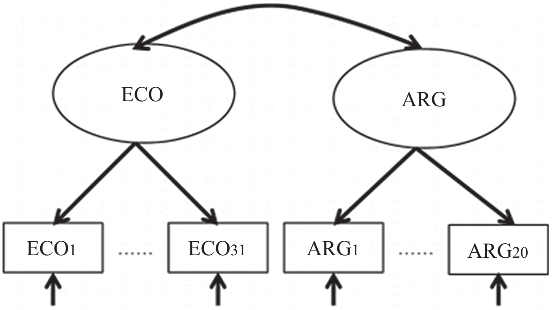


Figure 13.6: Graphical presentation of the second model, the multidimensional between-item model.

ally higher when compared to calibrating each subset consecutively with unidimensional models, and the resulting correlation coefficient is unattenuated.

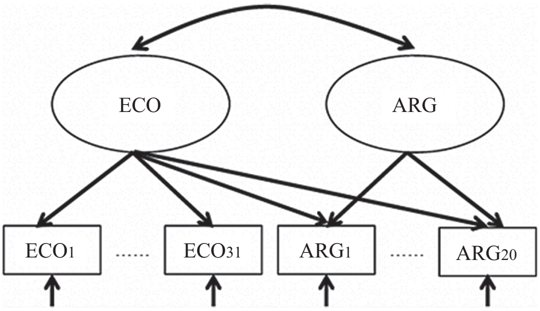


Figure 13.7: Graphical presentation of the third model, the multidimensional within-item model. Note that the two difficulties of the item parameters for the ARG$_k$ items represented by the arrows are constrained to be identical.

The multidimensional within-item model, shown in Figure 13.7, represents an alternative hypothesis (to the multidimensional between-item model) where the argumentation items are mapped to *both* the ecosystems and the argumentation constructs.

To investigate the underlying dimensionality of the science material, the *ConQuest* software (Adams et al., 2020) was used to fit the three models – the unidimensional model, the multidimensional between-item model, and the multidimensional within-item model. Because the unidimensional model can be considered to be nested within each of the two multidimensional models, the relative fit of the unidimensional and each multidimensional model can be evaluated using a likelihood ratio test, "which compares the difference in the deviance statistics between the two models with a chi-square test where the degrees of freedom is equal to the difference in the number of parameters estimated by each model" (Yao et al., 2015). The relationship between the two multidimensional models is not straightforwardly hierarchical (note that they have the same number of degrees of freedom), but the fits can still be compared using the Akaike information criterion (AIC; Yao & Schwarz, 2006) and Bayesian information criterion (BIC; Schwarz, 1978) indices.

13.4 Findings

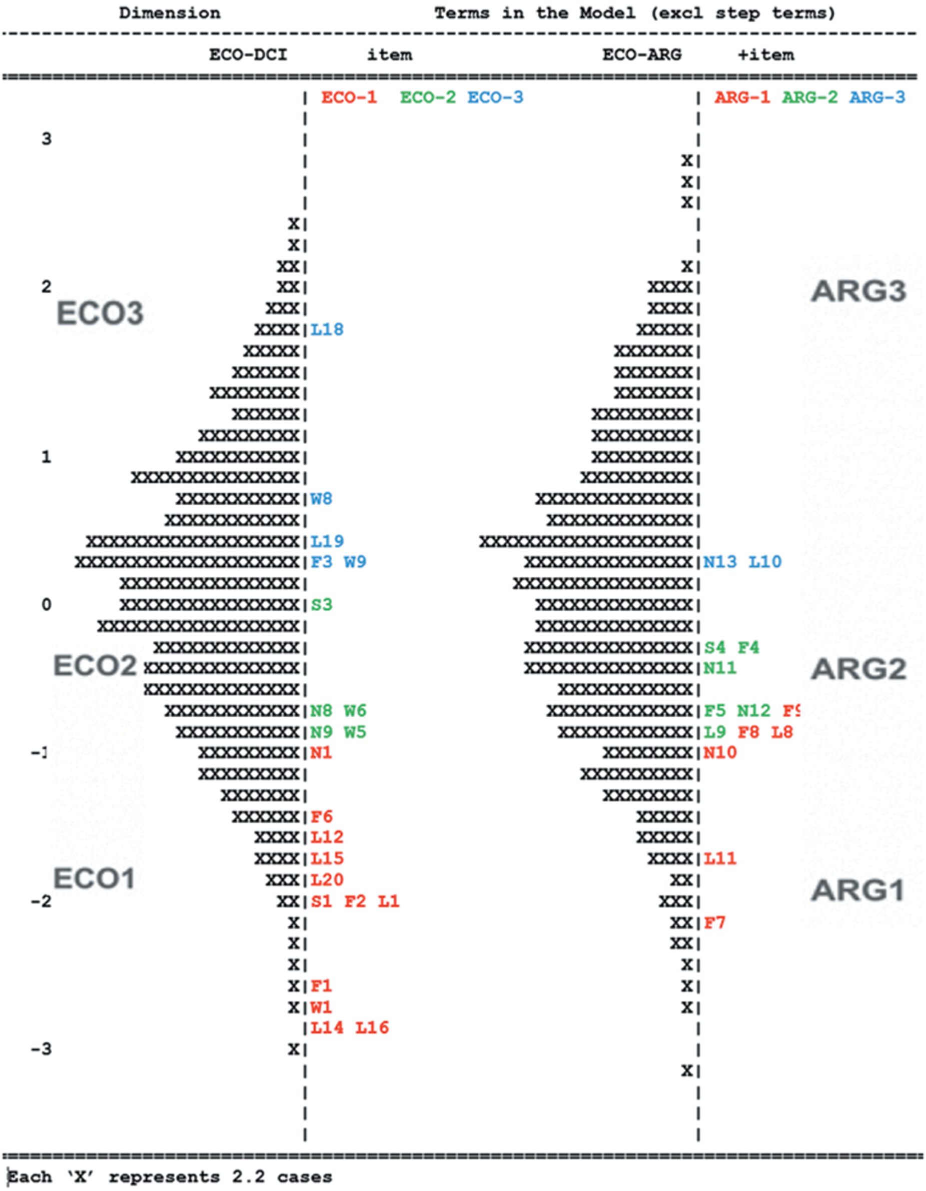
13.4.1 Internal structure validity evidence for the two constructs

The first step is to investigate whether the findings from the previous studies have been borne out in the new setting and new data collection. We will start here by examining the Wright maps from the two constructs.

A critically informative part of the output from a Rasch analysis is the Wright map, which is a graphical representation displaying the student ability estimates and item difficulty calibrations together on the same scale. Figure 13.8 shows the Wright maps generated for the multidimensional between-item model. The metrics of the two dimensions shown for the model have been transformed using a delta dimensional alignment technique (Feuerstahler & Wilson, 2021) to establish a common scale. This allows a direct comparison of the student abilities and item thresholds from the two dimensions (ecosystems and scientific argumentation).

As an orientation to Figure 13.8, the Wright map shows two columns of information, one for each construct, with ECO-DCI on the left and ARG-ECO on the right. Within each column the student ability distribution is on the left of the vertical line (representing the logit scale), and the item difficulty locations are shown on the right side of the vertical line. Each “X” in the student ability distributions represents 2.2 cases. The logit scale goes from lowest (at the bottom of the diagram) to highest (at the top of the diagram). Higher locations indicate higher ability waypoints on the construct for the students and more difficult items. The Wright map also shows how the students performed on the items. When a student’s estimated ability is at the same location as an item’s difficulty, the student has a 50% chance of answering that item in the correct way. This can be interpreted as the point of most active learning for the student. The items at the higher logit values would be more difficult for the student and the student would have less than a 50% chance of answering the question correctly. Conversely, if the item has a lower logit value, then the student would have more than a 50% chance of responding correctly to the item.

The Wright map provides valuable insights into the learning progressions. As shown in Figure 13.8, the range of item difficulties was reasonably well-matched to the range of student abilities in the sample across the two constructs, with sparser items at the upper end. Recall that both progressions identified four waypoints or waypoints of development to which the assessment items are mapped to, with Waypoint 1 being the first waypoint at which some understanding or competency is demonstrated. To facilitate reading, the



```
          Dimension                            Terms in the Model (excl step terms)
-----------------------------------------------------------------------------------
                                     ECO-DCI       item                  ECO-ARG      +item
===================================================================================
                          | ECO-1  ECO-2 ECO-3                |ARG-1 ARG-2 ARG-3
                          |                                   |
    3                     |                                   |
                          |                                  X|
                          |                                  X|
                          |                                  X|
                         X|                                   |
                         X|                                   |
                        XX|                                  X|
    2                   XX|                               XXXX|
                       XXX|                               XXXX|
                      XXXX|L18                           XXXXX|
                     XXXXX|                            XXXXXXX|
                    XXXXXX|                            XXXXXXX|
                  XXXXXXXX|                            XXXXXXX|
                    XXXXXX|                          XXXXXXXXX|
                 XXXXXXXXX|                          XXXXXXXXX|
    1          XXXXXXXXXXX|                          XXXXXXXXX|
           XXXXXXXXXXXXXXX|                         XXXXXXXXXX|
               XXXXXXXXXXX|W8                   XXXXXXXXXXXXXX|
              XXXXXXXXXXXX|                      XXXXXXXXXXXXX|
       XXXXXXXXXXXXXXXXXXX|L19             XXXXXXXXXXXXXXXXXXX|
      XXXXXXXXXXXXXXXXXXXX|F3 W9               XXXXXXXXXXXXXXX|N13 L10
          XXXXXXXXXXXXXXXX|                   XXXXXXXXXXXXXXXX|
    0     XXXXXXXXXXXXXXXX|S3                   XXXXXXXXXXXXXX|
        XXXXXXXXXXXXXXXXXX|                     XXXXXXXXXXXXXX|
             XXXXXXXXXXXXX|                    XXXXXXXXXXXXXXX|S4 F4
            XXXXXXXXXXXXXX|                    XXXXXXXXXXXXXXX|N11
            XXXXXXXXXXXXXX|                       XXXXXXXXXXXX|
              XXXXXXXXXXXX|N8 W6                 XXXXXXXXXXXXX|F5 N12 F9
               XXXXXXXXXXX|N9 W5                  XXXXXXXXXXXX|L9 F8 L8
   -1            XXXXXXXXX|N1                         XXXXXXXX|N10
                 XXXXXXXXX|                         XXXXXXXXXX|
                   XXXXXXX|                           XXXXXXXX|
                    XXXXXX|F6                            XXXXX|
                      XXXX|L12                           XXXXX|
                      XXXX|L15                            XXXX|L11
                       XXX|L20                              XX|
   -2                   XX|S1 F2 L1                        XXX|
                         X|                                 XX|F7
                         X|                                 XX|
                         X|                                  X|
                         X|F1                                X|
                         X|W1                                X|
                          |L14 L16                            |
   -3                    X|                                   |
                          |                                  X|
                          |                                   |
                          |                                   |
                          |                                   |
===================================================================================
Each 'X' represents 2.2 cases
```

Figure 13.8: Wright map of the multidimensional between-item model.

waypoints on the Wright map are color-coded, with items designed to target Waypoint 1 in red, items to target Waypoint 2 in green, and items to target Waypoint 3 (the highest waypoint) in blue. The items are also labeled based on the task and item information: for instance, “N10” tells us that the item at that location is Item 10 from the “Invasive” (N) task, which is targeted at Waypoint 1. One can see that the items shown in Figure 13.4 (N1 and N10) performed as anticipated at Waypoint 1.

We can see that for ECO (on the left-hand side of Figure 13.8), most of the items were found to be reasonably well-located within "bands" based on their hypothesized waypoints. Students were also reasonably well-distributed across the waypoints, with about half of the students located at Waypoint 3 and the remaining students roughly evenly distributed across Waypoints 1 and 2. At the lowest waypoint (Waypoint 1), students are at the point of learning about the relationships between two populations of organisms that are directly interacting with each other, such as predator-prey. At the intermediate waypoint (Waypoint 2), students are learning about more complex interrelationships including competition for abiotic resources and interactions between two populations that are not directly interacting with each other. Then, at the highest waypoint (Waypoint 3), students are learning about even more complex relationships between three or more populations, including microscopic organisms, which involve indirect interactions or competition for resources.

The story for ARG is quite analogous to that for ECO. Most of the items were found to be reasonably well-located within "bands" based on their hypothesized waypoints. Students were also reasonably well-distributed across the waypoints, with a bit less than half of the students located at Waypoint 3 and the remaining students roughly evenly distributed across Waypoints 1 and 2. At the lowest waypoint (Waypoint 1), students are at the point of learning about claims and evidence. At the intermediate waypoint (Waypoint 2), students are learning about how to link claims and evidence with warrants to make complete arguments. Then, at the highest waypoint (Waypoint 3), students are learning about comparing one argument with another.

Thus, examining both of the constructs, it can be seen that, in general, the locations of the items are quite reasonably consistent with the hypothesized construct maps. In addition, the Wright map suggests that students tend to progress along the two constructs quite consistently at about the same timing, as supported by how the bands in the two constructs line up in parallel.

13.4.2 Comparing the model fit

The second step is to explore evidence for the psychometric dimensionality of the constructs – this we will accomplish by examining the relative statistical fit of the different models (for statistical significance), and also by looking at the effect size (i.e., correlation between the constructs).

Table 13.2 shows the model fit statistics of the unidimensional, multidimensional between-item, and multidimensional within-item models. The difference in deviance between the multidimensional between-item model and unidimensional model is 7.29 with two degrees of freedom, which is statistically significant ($p < 0.05$). On the other hand, the difference in deviance between the multidimensional within-item model and unidimensional model is 3.07 with two degrees of freedom, which is not statistically significant (p = 0.11). In addition, note that for both the AIC and BIC indices, the between-item model is indicated to fit better than the within-item model. This implies that the multidimensional between-item model is the best fitting model. For this model, the *expected a posteriori* (EAP) person separation reliability is 0.73 for the content (ECO) dimension and 0.71 for the practice (ARG) dimension. This provides evidence that the argumentation items collectively measure a construct that is distinguishable from that measured by the content items.

Table 13.2: Summary of analysis of models.

Model	AIC	BIC	Deviance	Number of parameters
Unidimensional	44,492.87	44,765.09	44,388.87	52
Multidimensional between-item	44,489.58	44,772.27	44,381.58	54
Multidimensional within-item	44,493.80	44,776.49	44,385.80	54

In addition, the unattenuated correlation between the content and practice dimensions is 0.82, somewhat higher than the value found by Yao et al. (2015), but which is consistent with the expectation that students' competency in scientific argumentation should be positively correlated with their understanding of the science content.

The finding above is important because it answers the question of interest here, *what is the relationship between the content and practice (competency) constructs*? If the unidimensional or multidimensional within-item model had fitted the data better, then having the content disentangled from the practice would be much more challenging or even possibly not warranted. In other words, the results provide evidence that the argumentation items collectively measure a latent construct distinguishable from the construct measured by the content items and show that disentangling the content from the practice is possible.

13.5 Discussion and implications

Looking broadly, the correlation between the dimensions means that students with lower content (ecosystems) ability will likely have lower estimates in their practice (argumentation) competency contextualized in that content domain. Figure 13.8 shows that it is not just correlation involved here: students are succeeding at rates that show a *match* between the respective construct waypoints of argumentation and the scientific content of ecosystems, and reasonable progress is being made in both. The disclaimer here is that the supporting data comes from a cross-sectional rather than a longitudinal study.

The use of learning progressions as a resource in crafting assessments that coordinate with curriculum and pedagogy can be seen as beneficial. However, because of their speculative nature, it is critical to subject them to rigorous evaluation. In order to foster a beneficial influence, learning progressions need a robust theoretical framework and empirical verification. This paper shows a way to leverage existing high-quality learning progressions and associated items and build upon them for further exploration and empirical validation.

The empirical evidence from our analysis revealed a consistent pattern that lends support to the previous research (Dozier et al., 2023; Osborne et al., 2016). That is, the argumentation construct and the ecosystems construct waypoints were confirmed in this study as in the previous studies. One caveat is that the data used in this study included some of the data from Dozier et al.'s (2023) study.

This study builds on two previous studies about learning progressions for (a) student understanding of ecosystems (Dozier et al., 2023) and (b) scientific argumentation (Osborne et al., 2016), and extends them both by investigating their relationship. Indeed, the call for researchers to investigate the three dimensions of learning made by the NGSS has been partially answered here. By empirically investigating the two learning progressions together, we are able to see that they are related but distinct, which has implications for research and practice. This study leverages a previous methodological approach outlined by Yao and colleagues (2015) to understand and disentangle the relationship between the content and the practice. Additionally, this study extends Yao and colleagues' (2015) findings by applying their methodology in another science context (ecosystems).

Osborne and colleagues (2016) investigated scientific argumentation situated in general (familiar) scenarios in addition to within the physical science context of changes of state and compared the structure of both contexts. While they found a similar structure for argumentation in the science context and the familiar context, they found that arguing in the science context was more difficult than arguing in the familiar context and hinted that "there may be something particular about argumentation situated in scientific contexts that makes it more difficult than argumentation in familiar contexts" (p. 836). A possible future direction in this line of research could be to compare argumentation across multiple science domains to further investigate those relationships.

13.6 Final thoughts

Using the BAS methodology, this study extends and builds on studies about learning progressions in scientific practices (specifically, scientific argumentation) and science content (understanding of interdependent relationships in ecosystems). We find that the content and practice can indeed be aligned and disentangled with the use of theoretically robust learning progressions.

Extending Carol Weiss' (1980) ideas on *accretion*, we can begin to navigate a path forward for science assessment policy and the future use of results and methodology.

With the evidence presented here and elsewhere (Dozier et al., 2023; Osborne et al., 2016; Yao et al., 2015), we are beginning to see an accumulation of examples of how learning progressions can be validated, how dimensions can be aligned and disentangled, and how empirical approaches can be utilized to produce grounded results. However, more evidence is needed before policy recommendations can be made based on solid empirical study. For instance, the field needs valid learning progressions in earth science, more evidence about the relationships between science content and practice, and more evidence about the relationships among science content domains, scientific practices, and what the *Framework* (National Research Council, 2012) and the *Standards* (NGSS Lead States, 2013) define as CCCs such as *patterns*. These ideas need to be extended into the classroom as well.

This paper serves as an example from the field of education in support of a shared conceptual frame of reference for measurement across the sciences, as proposed by Mari et al. (2023). Clearly, significant conceptual and operational barriers will have to be overcome to achieve an expanded SI. Results such as those reported here suggest there is a rich potential for substantive productivity that could be multiplied many times over if communications were guided by a common frame such as a shared unit system.

Funding: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A160320 to the University of California, Berkeley. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

This material is based upon work supported by the National Science Foundation under Grant No. 2010322. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23.

Adams, R. J., Wu, M. L., Cloney, D., & Wilson, M. R. (2020). *ACER ConQuest: Generalized item response modelling software* [Computer software]. Version 5. Australian Council for Educational Research.

Alonzo, A. C., & Elby, A. (2019). Beyond empirical adequacy: Learning progressions as models and their value for teachers. *Cognition & Instruction*, *37*, 1–37.

American Association for the Advancement of Science. (2001). *Atlas of Science Literacy Vol. 1 and 2*. Washington, DC: National Academies Press.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32–42.

Cavagnetto, A. R. (2010). Argument to foster scientific literacy: A review of argument interventions in K–12 science contexts. *Review of Educational Research*, *80*(3), 336–371.

Chandler, M. J. (1992). The development of dynamic system reasoning. *Human Development*, *35*, 121–137.

Chi, S., Wang, Z., & Liu, X. (2023). Assessment of context-based Chemistry problem-solving skills: Test design and results from ninth-grade students. *Research in Science Education*, *53*, 295–318.

Dozier, S. J., MacPherson, A., Morell, L., Gochyyev, P., & Wilson, M. (2023). A learning progression for understanding interdependent relationships in ecosystems. *Sustainability*, *15*, 14212.

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education*, *47*(2), 123–182.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education*, *88*(6), 915–933.

Feuerstahler, L., & Wilson, M. (2021). Scale Alignment in the Between-Item Multidimensional Partial Credit Model. *Applied psychological measurement*, *45*(4), 268–282. https://doi.org/10.1177/01466216211013103

Kaldaras, L., Akaeze, H., & Krajcik, J. (2023). Developing and validating a Next Generation Science Standards-aligned construct map for chemical bonding from the energy and force perspective. *Journal of Research in Science Teaching*, *2023*, 1–38.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education*, *94*, 259–281.

Gotwals, A. W., Songer, N. B., & Bullard, L. (2012). Assessing students' progressing abilities to construct scientific explanations. In A. C. Alonzo & A. W. Gotwals (Eds.). *Learning progressions in Science: Current challenges and future directions* (pp. 183–210). Rotterdam, the Netherlands: SensePublishers.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. *Handbook of Educational Psychology*, *77*, 15–46.

Hayes, M. L., Plumley, C. L., Smith, P. S., & Esch, R. K. (2017). *A review of the research literature on teaching about interdependent relationships in ecosystems to elementary students*. Chapel Hill NC: Horizon Research, Inc.

He, P., Zhai, X., Shin, N., & Krajcik, J. (2023). Applying Rasch measurement to assess Knowledge-in-Use in science education. In X. Lui & W. Boone (Eds.). *Advances in applications of Rasch measurement in science education* (pp. 315–347). Cham, Switzerland: Springer Nature.

Hokayem, H., & Gotwals, A. W. (2016). Early elementary students' understanding of complex ecosystems: A learning progression approach. *Journal of Research in Science Teaching*, *53*, 1524–1545.

Jin, H., Mikeska, J. N., Hokayem, H., & Mavronikolas, E. (2019). Toward coherence in curriculum, instruction, and assessment: A review of learning progression literature. *Science Education*, *103*, 1206–1234.

Jin, H., Shin, H. J., Hokayem, H., Qureshi, F., & Jenkins, T. (2017). Secondary students' understanding of ecosystems: A learning progression approach. *International Journal of Science and Mathematics Education*, *17*, 217–235.

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement*, 2nd ed., (Springer series in measurement science and technology). Springer.

Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A construct-modeling approach to develop a learning progression of how students understand the structure of matter. *Journal of Research in Science Teaching*, *54*, 1024–1048.

National Assessment of Educational Progress (NAEP). (2019). *Science Framework for the 2019 National Assessment of Educational Progress*. US Department of Education; National Assessment Governing Board.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: National Academies Press.

Organization for Economic Co-operation and Development (OECD). (2012). *OECD science, technology and industry outlook 2012*. Paris, France: OECD.

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S.-Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, *53*(6), 821–846.

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika*, *53*, 349–359.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives*, *4*(1–2), 1–98.

Smith, C. L., Wiser, M., Anderson, C. W., Krajcik, J., & Coppola, B. (2004). *Implications of research on children's learning for assessment: Matter and atomic molecular theory. Paper commissioned by the Committee on Test Design for K-12 Science Achievement*. Washington, DC: National Research Council.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching*, *46*, 610–631.

Toulmin, S. (1958). *The uses of argument*. Cambridge, United Kingdom: Cambridge University Press.

Wang, W.-C., Wilson, M., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson & G. Engelhard (Eds.). *Objective measurement: Theory into practice* (Vol. IV, pp. 139–156). Norwood, NJ: Ablex.

Weiss, C. H. (1980). Knowledge creep and decision accretion. *Science Communication*, *1*(3), 381–202.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*(6), 716–730.

Wilson, M. (2023). *Constructing measures: An item response modeling approach*. (2nd ed.), New York, NY: Routledge.

Wilson, M., Scalise, K., & Gochyyev, P. (2019). Domain modeling for advanced learning environments: The BEAR Assessment System Software. *Educational Psychology*, *39*(10), 1199–1217.

Wiser, M., & Smith, C. L. (2008). Learning and teaching about matter in grades K-8: When should the atomic-molecular theory be introduced? In S. Vosniadou (Ed.). *International handbook of research on conceptual change* (pp. 205–239). Hillsdale, NJ: Erlbaum.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, *30*(6), 469–492.

Yao, S.-Y., Wilson, M., Henderson, J. B., & Osborne, J. (2015). Investigating the function of content and argumentation items in a science test: A multidimensional approach. *Journal of Applied Measurement*, *16*(2), 171–192.

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching*, *39*(1), 35–62.

Dhanya Nantha Kumar and H. B. Joshi

14 Patient-centered outcome assessments in surgical disciplines: an overview using example of the Urinary Stones and Intervention Quality of Life measure for kidney stone disease

Abstract: Surgical discipline involves rigorous assessments of outcomes, relevant to both surgeons and the recipients of care, over the short term and long term. The outcomes carry significance to other stakeholders such as the resource providers and industry partners. Patient-reported outcomes (PRO), which contribute to these assessments, are increasingly considered to be an important part of person-centered practices. In this chapter, we examine the current status of PRO assessments in the surgical field. Our focus is on applications of advanced measurement techniques and their adoption in processes relevant and useful to the various clinical stakeholders. We have divided the chapter into two parts. In the first part, we focus on the principles behind the development and validation of the Urinary Stones and Intervention Quality of Life scale, a disease- and intervention-specific PRO scale for urinary stone disease. We describe the framework, in which this new instrument has been developed, and we explore how probabilistic conjoint measurement theory has added scientific rigor to the traditional methods of PRO reliability and validity assessment. In the second part of the chapter, we provide a brief overview of the literature and examples of the probabilistic PRO measurement model's applications in the surgical branches of medicine. The current status of, and challenges surrounding, the development and application of PRO measurements of surgical outcomes are explored, with anticipations of the scope required for their wider adoption and applications, resulting in improved assessments.

Keywords: Outcome Assessments, Surgical Disciplines, USIQOL, Kidney Stones

14.1 Introduction

Patient-centered care is defined as being "respectful of and responsive to individual patient preferences, needs, and values, and ensuring that patient values guide all clinical decisions" (Committee on Quality of Health Care in America, 2001). It is a broad concept and forms one of the important components of quality care provision that in-

Dhanya Nantha Kumar, School of Medicine, Cardiff University, Cardiff, Wales, United Kingdom
H. B. Joshi, Department of Urology, University Hospital of Wales, Cardiff, Wales, United Kingdom

https://doi.org/10.1515/9783111036496-014

cludes person-centered approaches involving the total person and his/her own life, including caregivers. However, many factors are involved in the provision of good care, and this can be compromised by the lack of a body of well-developed theory, instruments, and evidence that substantiates the role and value of patient centered care in the broader medical context. Sometimes patient-centered care is mistakenly considered to contradict accepted standards of care. Providing and accounting for effective person-centered care must involve patients, using both qualitative and quantitative methods. These include tools such as interviews and direct observations, self-reported or performance-based measures, and more recently, measures from wearables and monitoring equipment (biomedical indicators).

Surgical disciplines manage many conditions that present in an acute and/or chronic form. The diversity of conditions demands a range of emergency and routine treatments, as well as one-off or repeated interventions. The decision to meet a patient's needs by opting for surgical interventions can pose risks. Many times, the disease can be treated in either a surgical or nonsurgical way with different risk-benefit ratios. The choice in favor of surgical management potentially involves an added degree of uncertainty due to the increased risk of morbidity and mortality during the perioperative period. Treatment options may present a clinical equipoise. Hence, patient-centered care, shared care decision making, and the understanding of patient preferences become very important.

14.1.1 Patient-reported outcome measurements (PROM): key to person-centered care

A patient-reported outcome measurement (PROM) is a report on patient's health condition that comes directly from the patient and plays an essential role in person-centered care (U.S Dept. of Health and Human Service Food and Drug administration, 2009). PROMs have been categorized as Generic or Disease-, Condition-, and Intervention-specific. Some generic measures are used as health economic tools to provide data on the quality-adjusted life years (QALYs). There can be some overlap between the aspects of health-related quality of life (HRQoL), measured by generic and disease-specific instruments. These tools have multiple applications.

In addition to the use of PROMs in randomized controlled trials to assess treatment effectiveness, there is growing interest in their use in routine HRQoL monitoring of patients and medical audits (Dept. of Health, 2010). Recent studies support the use of PROMs in clinical practice for improved shared decision-making and patient self-management (Kotronoulas et al., 2014). They have been found useful when there is a need to "identify triggers for surgery and potentially reduce the burden on services by limiting unnecessary or ineffectual procedures" (Kingsley & Patel, 2017). When used on a longitudinal basis, PROMs can track the progression and severity of disease and be incorporated as an adjunct to make changes to treatment and follow-up (Velikova et al., 2004).

There is evidence for the usefulness of the PROMs in clinical practice. PROMs facilitate the detection of physical or psychological problems (Bitton et al., 2014). PROMs compare favorably with other common clinical measures in terms of reliability (Snyder & Brundage, 2010). Many national surgical bodies advocate their use to evaluate outcomes, guide routine surgical practices, and in decision making. For example, the American Urological Association (AUA) guidelines state that treatment decisions about urinary calculi should incorporate patient preferences, influenced by HRQoL impacts, rather than being limited to clinical and radiological outcomes (Penniston & Nakada, 2016).

14.1.2 General considerations behind development and application of a PROM

PROMs would contribute more consistently to improving the evidence base, supporting patient-centered care, if the measurements were more solidly grounded in science and shown to be in accordance with international standards (US FDA and Scientific Advisory Committee, 2002). The ability of PROMs to improve decision-making depends on demonstrating how they accurately capture the burden of disease or effects of treatment. PROM data should clearly indicate the meaning of small changes to the scores and when there is a need to act or decide on management plans (Bitton et al., 2014).

The methodology for the development of a PROM was established over four decades ago and has continued to evolve. It involves a multiphase approach that includes construct definition: the qualitatively informed generation of items (questions). This is followed by pilot and field testing. The final instrument is expected to satisfy demands for reliability, validity, and responsiveness. Classical Test Theory (CTT) and its focus on ordinal scores formed the main basis for demonstrating measurement quality for many years, but it is now well recognized that measurements that comply with interval scaling requirements of conjoint additivity support higher quality inferences (Terwee et al., 2018).

Rasch, in 1960, proposed a theory of measurement, producing ratio/interval scales of both stimulus and object parameters (Rasch, 1960, 1961). Andrich (1988) stated that these models, relevant to the analysis of social science data, are the same as those of the laws of physics. Their perspective was further developed by other scientists focused on paired comparisons and has been more recently been said to provide “a specifically metrological approach to human-based measurement” (Fischer & Molenaar, 1995; Andrich & Marais, 2019; Linacre, 2000; Mari et al., 2023).

In measurements modeled to be conjointly additive, the probability of a specified response (e.g., right/wrong answer, or agreeable response) being a function of the difference between each individual person’s ability or performance, and the difficulty or challenge posed by each individual item. This is an approach to mathematical modeling where item values are calibrated and person abilities are measured on a shared continuum quantifying the latent trait. This approach cannot guarantee but supports the development of internally valid measurements that exhibit structural invariances, independent of the sample, with findings for samples extrapolating to population characteristics and clinically meaningful differences (Pendrill, 2014; Granger, 2007). This work underpins the current application of probabilistic measurement modeling in validations of contemporary PROMS.

Criteria for judging the quality of a PROM and its validity in the clinical field have been the subject of debate. For the application of PROMs in the clinical world, COSMIN guidelines were developed to evaluate the methodological quality of studies, intended to establish the measurement properties of HRQoL scales (Hobart & Cano, 2009). When selecting a robust PROM, these guidelines advocate the use of scales developed on the basis of probabilistic measurement modeling as this increases the likelihood of covering many important steps in validity assessments. These steps include the development of data fit to a model, the demonstration of unidimensionality and obtaining satisfactory discrimination as well as evaluative properties. These steps are discussed in the next section using the example of a disease- and intervention-specific PROM for urinary calculi.

14.2 Urinary Stones and Intervention Quality of Life (USIQoL) PROM: development and validation of a disease- and intervention-specific PROM for urinary calculi

Urolithiasis is a common condition that has a global incidence of 10% (prevalence range of 2–13% across continents) amongst the general population, with 50% of patients likely to form further stones within five years (Mokkink et al., 2010). The disease caused 550,000 emergency room visits in the USA in 2009 and over 30,800 hospital admissions in England in a single recent year (Pearle et al., 2005; Hospital Episodes Statistics Data, 2014). Stone patients miss an average of 47.9 h of work per year with additional hours lost due to ambulatory care visits (Bultitude & Rees, 2012).

There are different options for managing urinary calculi with expectant, medical or interventional treatments (Saigal et al., 2005), which can be multistage and carry different risks and success rates. Urolithiasis and its treatment(s) have an adverse effect on HRQoL and can compromise all areas of patient functioning (Türk et al., 2020; Raja et al., 2016). Attempts have been made to measure HRQoL of patients with urolithiasis (Penniston & Nakada, 2016). Generic measurement scales have been used for this, but often fail to capture the clinically relevant domains (Türk et al., 2020). This has led to the introduction of the new Urinary Stones and Intervention Quality of Life (USIQoL) questionnaire, a disease- and intervention-specific PROM that has been developed to meet the need for more relevant information.

The initial developmental work with patient interviews (62 patients and 30 family members) produced a conceptual framework and an initial long draft of the questionnaire. This generated 106 themes and 10 broad headings. These were mapped to a conceptual framework with removal of duplications to create item sets. A five-point rating scale (“not at all” to “a lot”) was selected for the initial draft.

Given the five-point rating scale and the items that are reasonably on-target (such that the sample measurement mean is near the item calibration mean, and the measurement and calibration ranges of variation overlap, with no significant floor or ceiling effects), a sample size of 25 to 60 will give 99% confidence that the item estimates are within 0.5 logits of their stable value (Patel et al., 2017). A sample ranging between 200 and 400 or 500 ought then to provide four or five class intervals. The validation was performed in 2 field tests and the analysis (polytomous extended response category, partial credit model) was performed using RUMM 2030 software.

14.2.1 Field test 1

Of the total sample of 250 patients, 212 participated in this phase. The revised version of the questionnaire included 60 items. It evaluated pain using different formats for rating the frequency of mild to unbearable pain, the intensity of the worst pain, day to day as well as average pain, etc. in 10 items overall; and also addressed physical and social health (including sex life, 18 items), psychological health (6 items), work performance (8 items) and travel/holiday

Table 14.1: Example of changes to the response categories due to disordered thresholds.

Initial draft:						
Since your current stone problems began, how much have you:	Not at all	A little	Quite a bit	Very much	A lot	N/A
Had difficulty sleeping?						
Felt depressed?						
Since your current stone problems began, have your stones made you reluctant about:	Not at all	A little	Quite a bit	Very much	A lot	N/A
Making a long journey?						
Planning a holiday because you might need to use unplanned medical services?						

Table 14.1 (continued)

Final draft:					
Since your current urinary stone problems began, **how much have you**	Not at all	A little	Quite a bit	A lot	N/A
Q7. Had difficulty sleeping? Q8. Felt depressed?					
Since your current, urinary stone problems, **have your symptoms made you reluctant about:**	Not at all/a little		Quite a bit/a lot		N/A
Q10. Making a long journey?					

issues (3 items). Fourteen items addressed additional problems, including those arising from treatments, and others involving help from the healthcare team and family members. Finally, a single global health question was included.

The results of a traditional analysis (classical test theory) for consistency and validity showed this draft of the USIQoL to be a reliable and valid measure of impact of stones on different domains. Reliability was satisfactory, given the diagnostic purpose of the scale [alpha: total scale (0.9), subscales (0.6–0.9)]. The corrected item total (0.3–0.8) and inter-item (0.4–0.9) correlations were satisfactory. Preliminary analyses of criterion validity were as expected (correlations with generic measures, range 0.3–0.8), demonstrating satisfactory early item-level validity.

Further measurement scaling analyses using conjoint additivity demonstrated many limitations that were not identified by the traditional (CTT) analysis. All scales indicated good to excellent reliability, with person separation indexes (PSI) ranging between 0.62 and 0.89, given the demands for precision tolerances imposed by screening and diagnostic applications. However, almost all scales had over 60% of the items with disordered thresholds (difficulty in distinguishing between responses “quite a bit” and “very much”), necessitating change from 5 to 4 or even 2 response categories (e.g., questions evaluating ability to travel for social reasons and leisure) (Table 14.1). This controversial step of collapsing adjacent categories was taken as a preliminary and provisional effort at creating a tool, meaningful to patients (Linacre, 1994; Adams et al., 2012).

In principle, the thresholds in any scale should demonstrate response categories, representing consistently increasing levels of the construct being measured (the correct ordering of the response categories is reflected in successive thresholds). We have observed that during clinical use, having thresholds that correspond to relatable ranges in the measured construct helps in improved understanding and patient acceptability of the scale items. Item fit is evaluated using the chi-square statistics to assess that the central property of item invariance (the hierarchical ordering of the items) does not vary across the trait measured. Fit residuals demonstrate the differences between the observed and expected data for each person and item. Each scale had items with significant fit residuals (12–60%), and residual correlations (50–90%), indicating redundancy of several items.

The removal of off-construct items, which provoked high-residual inconsistent responses, was conducted in an iterative manner, with the removal of a single item at a time, followed by reanalysis and creation of revised versions. It is important to note that this phase involves significant contributions from the clinicians and health care professionals who are experienced in the management of the target patient population. The statistical tests often result in an undecidable equipoise regarding item evaluations and so it is not always possible to select items based on analytic results alone.

The final item selection is always a multidisciplinary task. This is very important when the wider concept of validity of a PROM is to be considered. We found this to be helpful when subsequent application of the PROM in different clinical contexts was planned. The revised USIQoL included 19 questions sets divided as 5 scales of pain, social health (5 items each), physical, psychological health (4 each), and work (U.S Dept. of Health and Human Service Food and Drug administration, 2009) with 4 treatment items. This scale underwent a final validation study in a second field test.

14.2.2 Field test 2

In total, 369 of 390 patients participated in this phase. The analysis demonstrated that most of the items in the scales mapped out continua of increasing bother. The scales located items in a clinically sensible order with good sample match. Deviations from model expectations were marginal. Items excluded were pain (life interference, average and mild pain), social (sex, social life, and holiday), psychological (worry about kidney failing), and treatment (diet and device). The two treatment items (medication, water intake) were combined with the social scale. This transformed the USIQoL into a final 15-item measure.

We found that a revised scaling was necessary as items had superior fits when the 5-scale structure was changed to 3-scale, combining pain and physical health domains (PPH 6 items), psychological and social health domains (PSH 7 items), and work domain (2 items). Figures 14.1–14.3 illustrate satisfactory item-threshold distribution maps of subscales. Differential item functioning (DIF) evaluates the extent to which different groups within the sample (e.g., age, anatomical site of stone [kidney or ureter], and type of intervention). This is very important clinically, especially when the target population can be very heterogeneous. The stone disease has certain clinical features (ureteric vs renal stone, with or without underlying metabolic abnormality, etc.) that can carry different QoL impact for the groups, and influence management. We evaluated all 15 questions and 3 scales against different patient subpopulations, confirming adequate performance across sample groups. This is important in the context of its wider clinical application, where valid prediction of differential behaviors across patient subpopulations is essential.

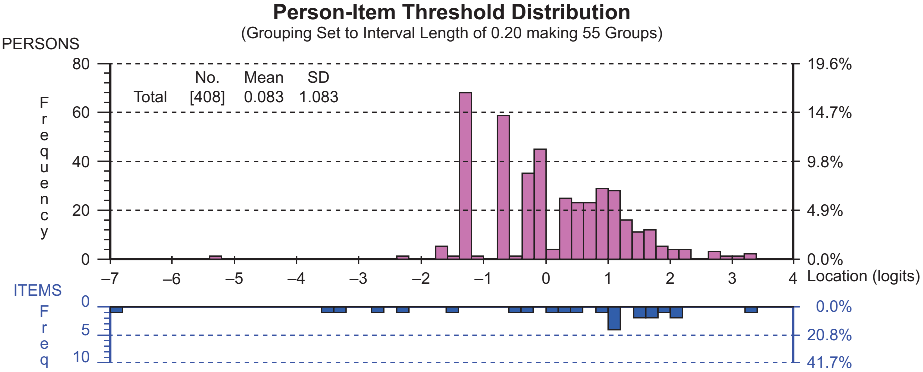


Figure 14.1: USIQoL: person–item threshold distribution – domain PPH.

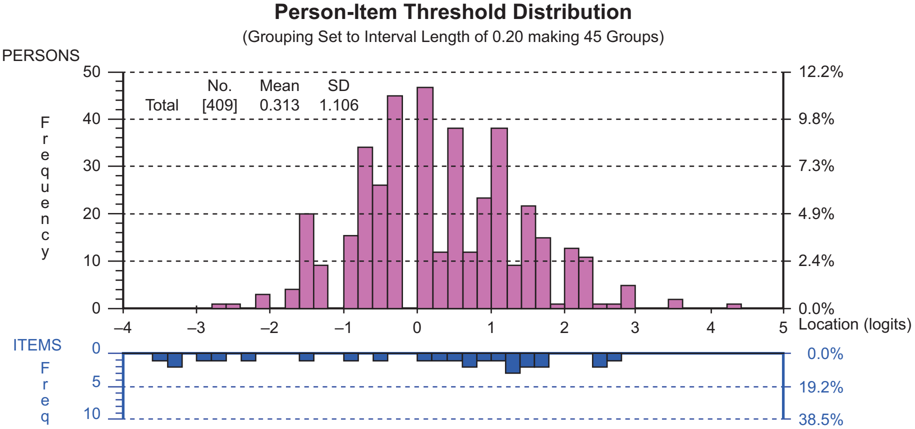


Figure 14.2: USIQoL: person–item threshold distribution – domain PSH.

Unidimensionality evaluations determine if any identifiable constructs are exhibited in the data after the main dimension has been considered. Model fit statistics indicate that all three scales of the USIQoL showed satisfactory unidimensionality. Pain, along with physical symptoms, which drives most of the clinical assessments, has clear impact. Pain, being the most complex construct to assess, was tested extensively before finalizing its format. Similarly, issues regarding work pose special data consistency problems because they are important to all stakeholders but not applicable to all patients. Conversely, the psychosocial scale is likely to be a good indicator of issues not evaluated routinely in clinical practice, and also of the longer-term impact of the condition, which could drive treatment choices. The USIQoL captures all of these dimensions well with the results quantified in a consistently interpretable, stable frame of reference.

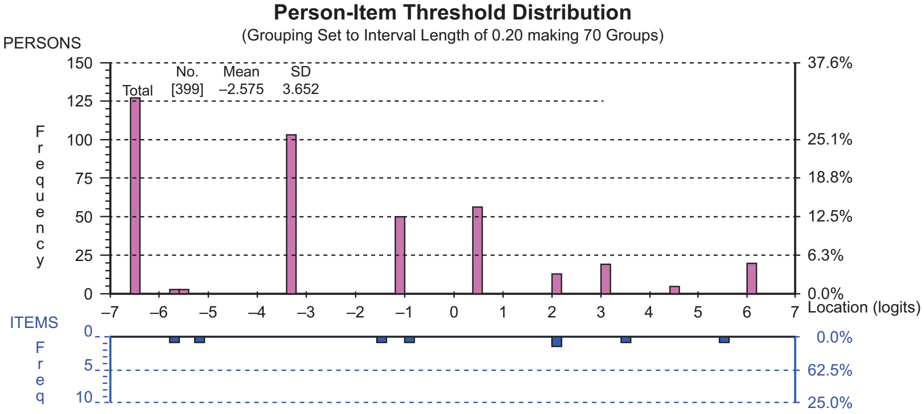


Figure 14.3: USIQoL: person–item threshold distribution – domain work.

The final USIQoL (3 scales and 15 items) is intended for self-administration, where patients rate the amount of bother attributed on a 4-point (1 = not at all, 2 = a little, 3 = quite a bit, or 4 = a lot) (Table 14.2). Scale scores are generated by summing items and transferring to a 0–100 (logit) scale, with high scores indicating greater patient bother. It provides an internally valid measurement, demonstrably invariant, independent of the sample and with findings extrapolating to population measures of clinically meaningful differences. The final item selection in USIQoL was based on the appraisals of the analyses against clinical relevance and measurement criteria. Psychometric evaluation showed that all three scales satisfied criteria for acceptability, validity, and reliability. The logit scoring for each scale offers different scores, allowing clearer identification of the impact across different domains. The results from traditional validity assessments alone suggested that the long draft of the USIQoL satisfied most of the criteria, until probabilistic measurement demonstrated many targeting problems (e.g., disordered responses and item redundancies). This highlighted the value of conjoint probabilistic measurement to conduct item-level analyses that guide precise item selection, and rectify problems with scales.

14.2.3 Clinical application of the USIQoL: establishing validity in a wider context

It has been suggested that although robust psychometric properties of a PROM, based on consensus statements, are a precondition to use, a PROM's validity in fact lies in the sound argument that a network of empirical evidence supports the intended interpretation in a particular context (Andrich, 2013). This idea was explored by conducting a feasibility study to see if the USIQoL can be used as an aid in outpatient settings to optimize the traditional follow-up of patients with urinary calculi. Most patients

Table 14.2: USIQoL final draft of the PROM.

Urinary Stones and Intervention Quality of Life Measure

The USI-QoL – Stone Disease©

We are interested in knowing how your quality of life has been affected by your **current urinary stone problems.**

Please answer all questions on the next page, in order, by ticking the appropriate boxes.

If your feel a question is not applicable to you, please tick the ‘N/A’ column.

Today’s date: _________________

Date of birth: _________________

We thank you for taking the time to complete this questionnaire

Please think about current **problems that are due to your urinary stones**

Since your current urinary stone problems, and due to urinary stone problems, **how much do you suffer with**	Not at all	A little	Quite a bit	A lot	N/A
Q1. Severe to unbearable pain?					
Q2. Pain triggered by physical activity?					
Q3. The feeling you need to pass urine urgently?					
Q4. Symptoms of a urinary tract infection (e.g. running temperature, feeling unwell and pain while passing urine)?					
Q5. Decreased or lack of appetite?					
Q6. Low energy?					
Since your current, urinary stone problems, **how much have you**	Not at all	A little	Quite a bit	A lot	N/A
Q7. Had difficulty sleeping?					
Q8. Felt depressed?					
Since your current, urinary stone problems, **with regards to the future, how much are you worried about:**					
Q9. More symptoms from your stones in the future?					
Since your current, urinary stone problems, **have your symptoms made you reluctant about:**	Not at all /A little		Quite a bit/A lot		N/A
Q10. Making a long journey?					
Since your current urinary stone problems, **how much have you had to visit the following, due to your symptoms:**	Not at all	A little	Quite a bit	A lot	N/A
Q11. GP or hospital during normal working hours?					

Table 14.2 (continued)

Urinary Stones and Intervention Quality of Life Measure					
The USI-QoL – Stone Disease©					
Since your current urinary stone problems, **how much have you found yourself having problems with:**					
Q12. Having to take medication (painkillers, preventative treatment etc.)?					
Q13. Increasing your water intake?					
Work					
Please mark 'Not applicable' (N/A) if currently not working (paid employment).					
Since your current urinary stone problems **with regard to your job, how much:**	Not at all	A little	Quite a bit	A lot	N/A
Q14. Have you needed to take time off work?					
Q15. Has your stone disease interfered with your ability to do your job?					

with urolithiasis undergo long-term follow-up involving regular clinic review and imaging to prevent or identify possible complications early. This is resource-intensive, involves exposure to ionizing radiation, and is not without diagnostic limitations. Furthermore, there are wide variations in practices. Deciding the optimal frequency and duration of follow-up for stones is a longstanding problem with little evidence base and alternatives. The National Institute of Clinical Excellence (NICE) in the UK indicated that currently no recommendations can be made regarding follow-up and that more research is needed (Hawkins et al., 2018).

The important question in need of answering when a patient with urolithiasis attends a clinic is whether the stone(s) need an intervention to treat or can be monitored. To this end, it would be important to know if the adoption of the USIQoL as a monitoring tool can assist clinical -making. This would be suitable if the results correlate well with those of traditional follow-up methods, or with outpatient review involving consultation and imaging. In the latest Urology Outpatient Transformation guide in the UK, "personalized follow up – patient-initiated follow-up" and "using remote monitoring" were highlighted as two key components within the scope of improved PROM-based follow-up (National Institute for Health and Care Excellence, 2023). Following the COVID pandemic, there are pressures for changes to outpatient practices and increased acceptance of alternative methods of follow-up (National Institute for Health and Care Excellence, 2023).

Hence a feasibility study was conducted to establish the validity of the USIQoL in a wider clinical context (Getting It Right First Time (GIRFT), 2023), with three objectives:

1) To assess the validity of the USIQoL as an outcome measure in the outpatient setting by establishing its correlation with the traditional follow-up (current standard of care).
2) To develop valid USIQoL cutoff scores that can reliably differentiate between patients who need active treatment against those who do not and facilitate a follow-up strategy, including remote methods.
3) To define the Minimal Clinically Important Difference (MCID) for the USIQoL, defined as the minimal change in the score, considered to be relevant by patients and physicians.

Initially, the USIQoL-based decision model was developed using existing data. Subsequently, a prospective, single-blind validation of the model for outpatients was conducted. For subjective measures, in general, including the application of the PROMs, the FDA recommends different types of anchors as external criteria, approximating truth, to generate relevant thresholds for meaningful within-patient change. These recommended anchors are

1) well-established clinical outcomes (intervention or not in our case);
2) global impressions of change in stone-related symptoms; and
3) static – current-state global impression of severity (EQ-5D PROM, in this case).

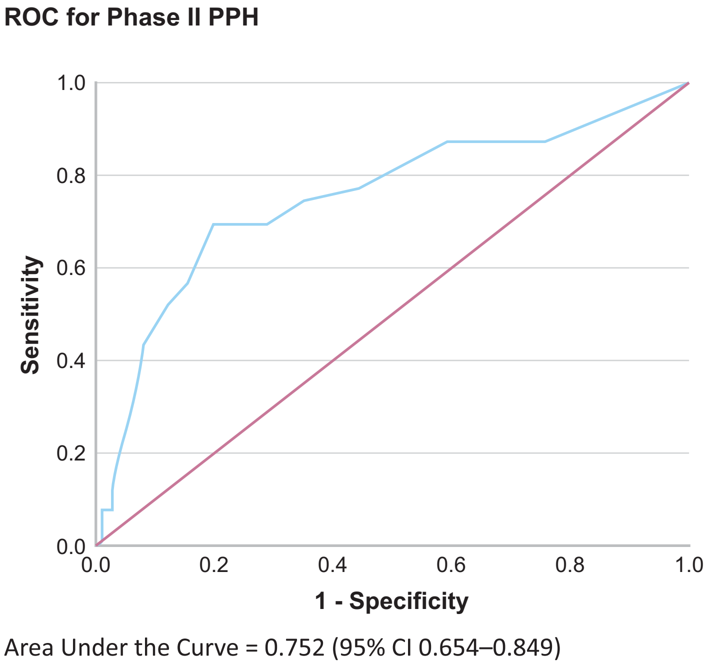
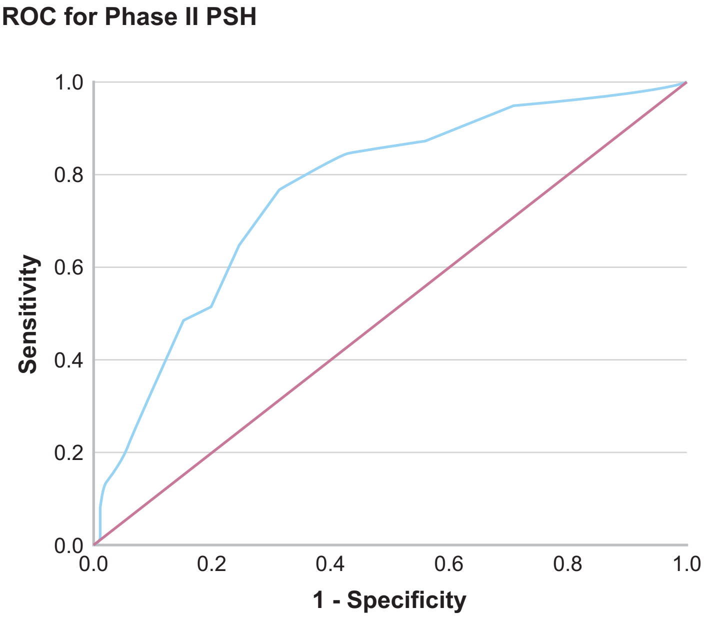


Figure 14.4: ROC curves for PPH (Pain and Physical health) domain – Phase II.

For the purposes of this study, USIQoL measurements from the two major domain scales (PPH and PSH) were considered. The study assessed correlation between the USIQoL measurements and the outcomes listed above. The study helped to validate USIQoL cutoff standards to discriminate between patients' needs to intervene or not. Analysis involved binomial logistic regression (BLR) and receiver operating characteristic (ROC) curves.

Data from 455 patients showed that the relationship between USIQoL scores (Pain and Physical Health, PPH and Psycho-Social Health, PSH domains) and clinical outcomes were statistically significant [estimated odds: PPH 1.24, $p < 0.001$, 95% CI 1.13–1.36; PSH 1.179, $p < 0.001$, 95% CI 1.12–1.33]. The ROC values were >0.75 when an Area under curve (AUC) of 0.7 to 0.8 is considered acceptable (Jarvis et al., 2023) (Figures 14.4 and 14.5). This demonstrated satisfactory ability of the model to differentiate between the two clinical outcomes. The optimum cutoff measurements were found to be 9 (PPH) and 11 (PSH), based on the Youden index.


ROC for Phase II PSH

Area Under the Curve = 0.763 (95% CI 0.677–0.849)

Figure 14.5: ROC curve for PSH (psycho-social health) domain – Phase II.

There is a significant clinical interest in defining the MCID for a given PROM so that the magnitude of the clinical impact, or change, can be understood and standardized. It is well known that MCID is a complex concept with multiple facets and variable results, based on the methods used. Combinations of anchor- and distribution-based methods were used to give the best estimates. The Minimally Clinically Important Difference (MCID) for the domains scores was 3–4 points. The model demonstrated satisfactory sensitivity (0.90) and specificity (0.46). Using this, it was clear that the odds of patients expressing symptoms and then needing full clinical evaluation with imaging and active intervention, increased with the increasing USIQoL scores. The results confirmed good correlation and one-dimensionality between the PPH and PSH domains.

Thus, the feasibility study demonstrated good correlation between the PROM and the clinical outcomes, making it a valid aid for outpatients. The cutoff scores identify patients at risk. It provides a reliable tool for patient-centric evaluation and an alternative to the long-term traditional follow-up policy, and established validity of USIQoL in a wider context.

14.3 PROMs in surgical disciplines: overview of the literature

We explored the current status of the PROMs in surgical disciplines, with a focus on application of the metrologically oriented measurement theory. Although the formal systematic review is out of the scope of this chapter, we have worked out the broad trends and key messages using examples from the literature. The implications are discussed in more detail in the subsequent section.

For the search, “patient reported outcomes,” “surgery,” “applications,” “outcome metrology,” “Rasch analysis,” “conjoint measurement,” and “decision making” were the key words used. The search, covering over 3 decades, resulted in over 18,000 articles with PROM and over 8,000 articles with Rasch key words. The results covered studies with significant heterogeneity. These largely reported on the developmental work on PROMs, or comparative studies, when applied to a cohort of patients in a single or multicenter study.

At the micro-level, PROMs facilitate the detection of clinical problems and adherence to treatments (Bitton et al., 2014). Real-time access to the PROM data helps clinicians prioritize topics for discussion at review and improves patient–clinician communication (Rasmussen et al., 2021). At the meso-level, PROM data can help in comparative effectiveness research and evaluation of the impact of interventions (Lavallee et al., 2016). There are four main mechanisms used internationally for the routine collection and aggregation of PRO information (Greenhalgh, 2009). Some of these have been exclusively used in the area of surgical practices:

A) Pre- and post-procedure data collection from patients undergoing selected elective surgeries to assess hospital performance (e.g., the National NHS PROMs program): Four surgical procedures were initially chosen to be included in the national PROMs program (2009 on), mandated in the NHS Outcomes, and included total hip replacement, total knee replacement, varicose veins (until 2017), and groin hernia surgery (until 2017). The main aim was to benchmark procedural outcomes across different trusts (Williams et al., 2016).

B) Computer-assisted testing using banks of questions that capture generic patient-reported outcomes, common across a number of chronic conditions (e.g., the US-based Patient Reported Outcomes Measurement Information System (PROMIS) initiative): This is aimed at providing patient-level data using pre-prepared question banks covering different domains (Coles, 2010).

C) Inclusion of PROMs within disease-specific clinical registries (e.g., the Swedish Healthcare Quality Registries).

D) International initiatives to develop standard outcome measurement sets, including PROMs, to foster international benchmarking (e.g., International Consortium for Health Outcomes Measurement).

This literature thus demonstrates the current wide-ranging applications of PROMs.

14.3.1 PROMS and evaluation using additive conjoint measurement modeling techniques

14.3.1.1 Development of new PROMS

Many new PROMs covering different surgical disciplines have been developed using conjoint measurement theory and modeling over the last 15 years, with many employed in evaluating clinical trial outcomes (PROMIS®; Joshi et al., 2022; Pesudovs et al., 2004). A new 20-item Quality of Life Impact of Refractive Correction (QIRC) questionnaire, which quantifies the QOL of people with refractive correction by spectacles, contact lenses, and refractive surgery in the prepresbyopic age group, was developed by Pesudovs et al. in 2004 and has been shown to have broad applicability for cross-sectional and outcomes research (Joshi et al., 2022). Similarly, the BREAST-Q is a PROM used to assess the unique outcomes of breast surgery patients that was developed in 2009 using conjoint measurement modeling; it is composed of three procedure-specific modules: augmentation, reduction, and reconstruction and has been used in multiple studies along with linguistic validations (Pesudovs et al., 2004).

14.3.1.2 Reevaluation of existing PROMs

Over the last two decades, the properties of existing PROMs have been developer-evaluated. Surgical disciplines such as orthopedics and ophthalmology have been at the forefront in these efforts. The results are mixed and have repeatedly demonstrated and substantiated the importance of adopting rigorous measurement modeling theory and practice. Many existing PROMS have been found to have problems with suboptimal targeting, item fit, disordered thresholds, and a lack of meaningful and interpretable unidimensionality. This has raised questions about the validity of the measurements and the results generated using these PROMs.

The National Eye Institute Refractive Error Quality of Life instrument (NEI-RQL-42) is a commonly used questionnaire that seeks to measure refractive error-related quality of life (QoL). In light of the results produced by conjoint measurement modeling, the authors stated that NEI-RQL-42 questionnaire is deficient for all psychometric properties tested and advise clinicians or researchers to consider other questionnaires that have been more rigorously developed to meet standard psychometric properties (Pusic et al., 2009). Another study was conducted in patients with prostate cancer, undergoing radical surgical treatment. The outcomes from the surgery were monitored using the patient-reported outcome measure: Symptom Tracking and Reporting tool (STAR) (Alinden et al., 2011). This tool has four domains, which investigates sexual function, urinary function, bowel function, and overall quality of life. The study showed that urinary and sexual function scales produced inconsistent observations, insufficient to the task of measurement. The study concluded that further evaluation needs to be carried out to determine the suitability of this PROM.

A study of Patient- and Parent-Reported Outcome Measures in the International Consortium for Health Outcomes Measurement Standard Set for Cleft Lip and Palate came to similar conclusions (Protopapa et al., 2020). The study concluded that the NOSE and COHIP-OSS questionnaires were inaccurate, and that the CLEFT-Q questionnaire did not cover facial function and speech domains sufficiently. The study concluded that the PROMs used for cleft care do not satisfy the need for quantitative measurements of the outcomes produced.

Re-evaluations have also been conducted for many short-form versions of existing questionnaires (Apon et al., 2021; Multanen et al., 2020). The reviews show that, in spite of many advances over four decades, it is still challenging to select reliable tools (Lundström & Pesudovs, 2009). Of the 315 generic and condition-specific PROMs published between the 1980s and 2019, the vast majority were related to musculoskeletal conditions, with other patient-related outcomes related to cancer, gastrointestinal, mental health, and many other conditions. Of the 315 studies identified, 270 (85.7%) had been used in subsequent studies, and 45 did not have any online evidence of applications, following validation.

14.3.2 Challenges in using PROMs in clinical practice

There are multiple challenges in the implementation of PROMs and these encompass different aspects of PROM usage. The challenges can be identified at different stages:

A. Development

1) Scientifically rigorous modeling is essential when developing measuring tools that are valid and reliable. In this regard, there has been considerable confusion within the scientific and the clinical communities regarding the viability of meaningful quantification, and the associated terminologies used in the field of PROM development and validation. This has had deleterious impacts on the interpretation and adoption of PROMs by clinical teams (Churruca et al., 2021; Derriennic et al., 2019; Hobart et al., 2007). Efforts undertaken by different agencies, such as COSMIN, are intended to standardize the nomenclature (Hobart et al., 2010).
2) Establishing metrological standards is essential for maximizing the value of the widespread use of a PROM. COSMIN standards for the validity of measurements include criteria that can be met only if the PROM has been developed using additive conjoint measurement modeling and so demonstrates validity in a wider context than that available using ordinal measurement methods (Churruca et al., 2021; Derriennic et al., 2019; Hobart et al., 2007, 2010; Prinsen et al., 2018; Hawkins et al., 2018; Snyder & Brundage, 2010; Fisher, 2023; Allen & Pak, 2023; Massof & Bradley, 2023). This is a desirable long-term strategy that needs to be endorsed by all stakeholders. The data from such work would establish a robust evidence base for patient-centered practices; with ongoing application of the insights of the new institutional economics, such standards may one day be legally enforced, with significant implications for health care markets (Snyder & Brundage, 2010).

B. Clinical applications

1) The selection of instruments appropriate for a given range of conditions and interventions can be challenging. There is a need for standardized assessments of the psychometric properties and validities of PROMs so that information provided is sensitive, relevant, and specific to various contexts. Provisionally resolving the tensions between standardization and personalization (Fisher, 2023; Allen & Pak, 2023; Massof & Bradley, 2023; Lipscomb et al., 2007; Mallinson, 2024) via meaningful scaling and individualized reporting is essential for generalized improvements in deciding the superiority or inferiority of surgical approaches or policies (Massof & Bradley, 2023).
2) There is a need to improve the comparability of PRO measurements and data across different healthcare settings, countries, and cultures, which poses challenging but not intractable problems (Lipscomb et al., 2007; Mallinson, 2024).
3) Patient and stakeholder engagement with diagnostic, treatment, and follow-up processes can be facilitated by improved measurement, as high-quality, actionable information provided confidentially via easy-to-use electronic interfaces may work to increase response rates in contexts involving the need to complete the PROMS on a repeated basis (for pre and post intervention assessments) (Massof & Bradley, 2023). Concerns expressed by clinicians have

included the time and effort involved in data collection and analysis, and the provision of adequate resources to collect the data and its analysis. Investment is required when establishing platforms for data collection and optimizing the flow and analysis of the data but may pay remarkable returns when systems are well-designed (Snyder & Brundage, 2010).

4) Data needs to be presented in forms usable to all the stakeholders at all levels (Snyder & Brundage, 2010) with clear information on what, if anything, small changes to the scores actually mean clinically (Fisher, 2023; Allen & Pak, 2023); anything less can risk leading to clinician disengagement (Bitton et al., 2014). Measurements should provide information on quality indicators; PROM data has been used to this effect in the UK in national audits covering the index orthopedic procedures (Williams et al., 2016).
5) Studies have documented limitations of existing PROMs without the application of probabilistic conjoint modeling (Pusic et al., 2009; Alinden et al., 2011; Protopapa et al., 2020). Fresh perspectives on standard setting are needed to revise the old measurement scales or develop new ones.

14.3.3 Opportunities for PROM applications

Surgical disciplines continue to evolve as minimally invasive and robotic techniques increasingly complement patient-centered practices. This offers opportunities for incorporating PROMs at every level of practice.

Micro: These are at the clinician–patient interactions level, where measurement reports are individualized to specific patients in the course of care, and to specific clinicians in the course of clinical management (Sul, 2024; Chien et al., 2009). There is evidence of benefits from these processes (Wright et al., 1980) as they contribute to improved patient counselling, ahead of interventions and the development of appropriate patient information leaflets. However, it is yet not established if the individual health status outcomes are consistently improved or not.

Meso: At the meso-level, PROMs are widely shown to be effective (Derriennic et al., 2019; Hobart et al., 2007; Fisher, 2023; Allen & Pak, 2023). This applies to the comparative effectiveness research used to investigate benefits of different treatment and surgical interventions. PRO data used in the registries helped quality improvement programs (NHS UK PROMS programs) and has improved understanding of the variations in care, costs, and outcomes. One of the major applications of PROMs is in the adoption of Value-Based Health Care (VBHC) (Chien et al., 2018). Although healthcare funding varies between different settings worldwide, there is a gradual shift from fee-for-service to the more VBHC. It aims to reduce unnecessary variations and costs in the practices. Person-centered data would provide valuable support to such programs.

Macro: There is growing interest in the development of predictive theories and explanatory models capable of independently validating the construct measured (Squitieri et al., 2017; Melin et al., 2021, 2023). Work in this area and in the programs advanced by groups such as the International Consortium for Health Outcomes (ICHOM) will help to foster international benchmarking.

More work needs to be done to advance the adoption of mass-customizable PROMs in a uniform and structured fashion. Use of measurements based on probabilistic, additive conjoint modeling analysis, with well-defined MCIDs and validity in the wider contexts, can plausibly be expected to result in new levels of utility, effectiveness, and efficiency. Standards will need to be established for unit definitions, laboratory accreditation, conformity assessment, and quality-assured traceability (Chan et al., 2015). National and international standards bodies and specialty organizations will need to focus complex cross-disciplinary initiatives on the demands of practice to devise and set the necessary guidelines. Health care insurers, funders, providers, regulators, and advocacy groups will need to collaborate in new ways to provide the necessary support and infrastructure.

Clear and interpretable standards of these kinds will support the creation of an entirely new class of quality improvement programs. It will offer opportunities for the development of systems capable of guiding systematic responses to PROMS feedback. Improvements to health information systems and technology (Jeckelmann et al., 2023) will address barriers to data collection and workload management by implementing computer-adaptive and AI measurement strategies and integrating PROM data in health records (National Institute for Health and Care Excellence, 2023). Close attention to envisioning, planning, and resourcing the needed broad scope for training and professionally developing clinicians and associated staff will pay significant substantive and financial returns as we achieve the timely dissemination of more relevant and meaningful information.

References

Adams, R. J., Wu, M., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, *72*(4), 547–573.

Alinden, C. M., Skiadaresi, E., Moore, J., & Pesudovs, K. (2011). Subscale Assessment of the NEI-RQL-42 Questionnaire with Rasch Analysis. *Clinical and Epidemiologic Research*, *52*, 8.

Allen, D. D., & Pak, S. (2023). Improving clinical practice with person-centered outcome measurement. In W. P. Fisher Jr. & S. J. Cano (Eds.). *Person centered outcome metrology* (pp. 53–105). Springer.

Andrich, D. (1988). *Sage university paper series on quantitative applications in the social sciences*. Vol. series no. 07–068, Rasch models for measurement. Sage Publications.

Andrich, D. (2013). An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any ‘threshold disorder controversy. *Educational and Psychological Measurement*, *73*(1), 78–124.

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social, and health sciences*. Springer.

Apon, I., van Leeuwen, N., Allori, A. C., Rogers-Vizena, C. R., Koudstaal, M. J., Wolvius, E. B., Cano, S. J., Klassen, A. F., & Versnel, S. L. (2021). Rasch analysis of patient- and parent-reported outcome measures in the international consortium for health outcomes measurement standard set for cleft lip and palate. *Value in Health*, *24*(3), 404–412. doi:10.1016/j.jval.2020.10.019

Bitton, A., Onega, T., Tosteson, A. N. A., & Haas, J. S. (2014). Toward a better understanding of patient-reported outcomes in clinical practice. *American Journal of Managed Care*, *20*(4), 281–283. Accessed March 2, 2023, https://pubmed.ncbi.nlm.nih.gov/24884859/

Bultitude, M., & Rees, J. (2012). Management of renal colic. *BMJ*, *345*, e5499.

Chan, T. L., Perlmutter, M. S., Andrews, M., Sunness, J. S., Goldstein, J. E., & Massof, R. W., Low Vision Research Network (LOVRNET) Study Group. (2015). Equating visual function scales to facilitate reporting of Medicare functional g-code severity/complexity modifiers for low-vision patients. *Archives of Physical Medicine and Rehabilitation*, *96*(10), 1859–1865.

Chien, T. W., Chang, Y., Wen, K. S., & Uen, Y. H. (2018). Using graphical representations to enhance the quality-of-care for colorectal cancer patients. *European Journal of Cancer Care*, *27*(1), e12591.

Chien, T.-W., Wang, W.-C., Wang, H.-Y., & Lin, H.-J. (2009). Online assessment of patients' views on hospital performances using Rasch model's KIDMAP diagram. *BMC Health Services Research*, *9*, 135.

Churruca, K., et al. (2021). Patient-reported outcome measures (PROMs): A review of generic and condition-specific measures and a discussion of trends and issues. *Health Expect*, *24*(4), 1015–1024. doi:10.1111/hex.13254

Coles, J. (2010). PROMs risk adjustment methodology guide for general surgery and orthopaedic procedures. Hertfordshire, UK: Northgate Information Solutions (UK) Limited, 54 pp.

Committee on Quality of Health Care in America, (2001). Improving the 21st-century health care system. In Crossing the quality chasm: A new health system for the 21st century (ed., pp. 39–60). Washington, DC: National Academy Press.

Dept of Health. (2010). *Equity and Excellence: Liberating the NHS*. London Dept of health.

Derriennic, J., Nabbe, P., Barais, M., Lalande, S., Le Goff, D., Pourtau, T., Penpennic, B., & Le Reste, J. Y. Quality of primary care from the patient's point of view. A systematic review. Preprint, Universite de Bretagne Occidentale Faculte de Medecine et Des Sciences de la Sante de Brest, 2019, September 9. https://doi.org/10.21203/rs.2.14226/v1

Fischer, G. H., & Molenaar, I. (1995). *Rasch models: Foundations, recent developments, and applications*. Springer-Verlag.

Fisher, W. P., Jr. (2023). Measurement systems, brilliant results, and brilliant processes in healthcare. In W. P. Fisher Jr. & S. Cano (Eds.). *Person-centered outcome metrology* (pp. 357–396). Springer.

Getting It Right First Time (GIRFT). (2023). Urology outpatient transformation: a practical guide to delivery. Accessed March 2. https://www.gettingitrightfirsttime.co.uk/wp-content/uploads/2022/01/Urology_2022-01-12_Guidance_Outpatient-transformation.pdf

Granger, C. (2007). Rasch Analysis is important to understand and use for measurement. *Rasch Measurement Transactions*, *21*(3), 1122–1123.

Greenhalgh, J. (2009). The applications of PROs in clinical practice: What are they, do they work, and why? *Quality of Life Research*, *18*(1), 115–123. doi:10.1007/s11136-008-9430-611

Hawkins, M., Elsworth, G., & Osborne, R. (2018). Application of validity theory and methodology to patient-reported outcome measures (PROMs): Building an argument for validity. *Quality of Life Research*, *27*, 1695–1710.

Hawkins, M., Elsworth, G., & Osborne, R. (2018). Application of validity theory and methodology to patient-reported outcome measures (PROMs): Building an argument for validity. *Quality of Life Research*, *27*, 1695–1710.

Hobart, J., & Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: The role of new psychometric methods. *Health Technol Assess*, *13*(12), 1–200.

Hobart, J. C., Cano, S. J., & Thompson, A. J. (2010). Effect sizes can be misleading: Is it time to change the way we measure change? *Journal of Neurology, Neurosurgery, & Psychiatry*, *81*, 1044–1048.

Hobart, J. C., Cano, S. J., Zajicek, J. P., & Thompson, A. J. (2007). Rating scales as outcome measures for clinical trials in neurology: Problems, solutions, and recommendations. *The Lancet Neurology*, *6*, 1094–1105.

Hospital Episodes Statistics. (2014). National Health Service, http://www.hscic.gov.uk/hes.

Jarvis, R., Pallman, P., & Joshi, H. (2023). Is remote follow-up using patient reported outcome measure (PROM) feasible in patients with urolithiasis? The results of the first prospective feasibility study using urinary stones and intervention quality of life (USIQoL) core measure. *European Urology*.

Jeckelmann, B., & Edelmaier, R. (2023). Metrological infrastructure. *De Gruyter Series in Measurement Sciences*. De Gruyter Oldenbourg.

Joshi, H. B., Johnson, H., Pietropaolo, A., Raja, A., Joyce, A. D., Somani, B., Philip, J., Biyani, C. S., & Pickles, T. (2022). Urinary stones and intervention quality of life (USIQoL): Development and validation of a new core universal patient-reported outcome measure for urinary calculi. *European Urology Focus*, *8*(1), 283–290. doi:10.1016/j.euf.2020.12.011, Epub 2021 Jan 8. PMID: 33423970

Kingsley, C., & Patel, S. (2017). Patient-reported outcome measures and patient-reported experience measures. *BJA Education*, *17*(4), 137–144. doi:10.1093/bjaed/mkw060

Kotronoulas, G., Kearney, N., Maguire, R., et al. (2014). What is the value of the routine use of patient-reported outcome measures toward improvement of patient outcomes, processes of care, and health service outcomes in cancer care? A systematic review of controlled trials. *Journal of Clinical Oncology, 32*(14), 1480–1501. doi:10.1200/JCO.2013.53.5948

Lavallee, D. C., Chenok, K. E., Love, R. M., et al. (2016). Incorporating patient-reported outcomes into health care to engage patients and enhance care. *Health Aff (Millwood)*, *35*(4), 575–582. doi:10.1377/hlthaff.2015.1362

Linacre, J. M. (1994). Sample size and item calibration [or person measure] stability. *Rasch Measurement Transactions*, *7*(4), 328.

Linacre, J. M. (2000). Almost the Zermelo model? *Rasch Measurement Transactions*, *14*(2), 754. http://www.rasch.org/rmt/rmt142k.htm

Lipscomb, J., Gotay, C. C., & Snyder, C. F. (2007). Patient-reported outcomes in cancer: A review of recent research and policy initiatives. *CA: A Cancer Journal for Clinicians*, *57*, 278–300.

Lundström, M., & Pesudovs, K. (2009). Catquest-9SF patient outcomes questionnaire: Nine-item short-form Rasch-scaled revision of the Catquest questionnaire. *Journal of Cataract & Refractive Surgery* , *35*(3), 504–513. doi:10.1016/j.jcrs.2008.11.038, PMID: 19251145.

Mallinson, T. (2024). Extending the justice-oriented, anti-racist framework for validity testing to the application of measurement theory in re(developing) rehabilitation assessments. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*. De Gruyter.

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement*. 2nd ed, Springer Series in Measurement Science and Technology, Springer.

Massof, R. W., & Bradley, C. (2023). An adaptive strategy for measuring patient-reported outcomes. In W. P. Fisher Jr. & S. J. Cano (Eds.). *Person-centered outcome metrology* (pp. 107–150). Springer.

Melin, J., Cano, S., & Pendrill, L. (2021). The role of entropy in construct specification equations (CSE) to improve the validity of memory tests. *Entropy*, *23*(2), 212.

Melin, J., Cano, S. J., Gillman, A., Marquis, S., Flöel, A., Göschel, L., & Pendrill, L. R. (2023). Traceability and comparability through crosswalks with the NeuroMET memory metric. *Scientific Reports*, *13*(1), 5179.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & De Vet, H. C. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research*, *19*(4), 539–549. doi:10.1007/s11136-010-9606-8, Epub 2010 Feb 19. PMID: 20169472; PMCID: PMC2852520.

Multanen, J., Ylinen, J., Karjalainen, T., et al. (2020). Structural validity of the Boston Carpal Tunnel Questionnaire and its short version, the 6-Item CTS symptoms scale: A Rasch analysis one year after surgery. *BMC Musculoskeletal Disorders*, *21*, 609. https://doi.org/10.1186/s12891-020-03626-2

National Institute for Health and Care Excellence. (2023). Renal and ureteric stones: assessment and management (NG 118). Accessed March 2, https://www.nice.org.uk/guidance/NG118.

Patel, N., Brown, R. D., Sarkissian, C., De, S., & Monga, M. (2017). Quality of life and urolithiasis: The patient – Reported outcomes measurement information system (PROMIS). *International Brazilian Journal of Urology*, *43*, 880–886.

Pearle, M. S., Calhoun, E. A., & Curhan, G. (2005). Urologic diseases in America project: Urolithiasis. *The Journal of Urology*, *173*, 848.

Pendrill, L. R. (2014). Man as a measurement instrument [Special Feature]. *NCSLi Measure: The Journal of Measurement Science*, *9*(4), 22–33.

Penniston, K. L., & Nakada, S. Y. (2016). Treatment expectations and health-related quality of life in stone formers. *Current Opinion in Urology*, *26*, 50–55.

Penniston, K. L., & Nakada, S. Y. (2016). Treatment expectations and health-related quality of life in stone formers. *Current Opinion in Urology*, *26*, 50–55.

Pesudovs, K., Garamendi, E., & Elliott, D. B. (2004 Oct). The Quality of life impact of refractive correction (QIRC) questionnaire: Development and validation. *Optometry and Vision Science*, *81*(10), 769–777. doi:10.1097/00006324-200410000-00009, PMID: 15557851.

Prinsen, C. A., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., De Vet, H. C., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, *27*, 1147–1157.

PROMIS® (Patient-Reported Outcomes Measurement Information System®): US Department of health and human services: https://www.healthmeasures.net/explore-measurement-systems/promis

Protopapa, E., van der Meulen, J., Moore, C. M., & Smith, S. C. (2020). Assessment of a patient-reported outcome measure in men with prostate cancer who had radical surgery: A Rasch analysis. BMJ Open, 10(11), e035436. doi:10.1136/bmjopen-2019-035436

Pusic, A. L., Klassen, A. F., Scott, A. M., Klok, J. A., Cordeiro, P. G., & Cano, S. J. (2009). Development of a new patient-reported outcome measure for breast surgery: The BREAST-Q. *Plastic and Reconstructive Surgery*, *124*(2), 345–353. 19644246.

Raja, A., Hekmati, Z., & Joshi, H. B. (2016). How do urinary calculi influence health-related quality of life and patient treatment preference: A systematic review. *Journal of Endourology*, *30*, 727–743.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Danmarks Paedogogiske Institut. Danmarks Paedogogiske Institut.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability: Volume IV: Contributions to biology and problems of medicine* (pp. 321–333). http://www.rasch.org/memo1960.pdf. University of California Press

Rasmussen, A. A., Wiggers, H., Jensen, M., et al. (2021). Patient-reported outcomes and medication adherence in patients with heart failure. *European Heart Journal –Cardiovascular Pharmacotherapy*, *7*(4), 287–295. doi:10.1093/ehjcvp/pvaa097

Saigal, C. S., Joyce, G., Timilsina, A. R., et al (2005). Direct and indirect costs of nephrolithiasis in an employed population: Opportunity for disease management? *Kidney International*, *68*, 1808–14.

Snyder, C., & Brundage, M. (2010). Integrating patient-reported outcomes in healthcare policy, research and practice. *Expert Review of Pharmacoeconomics and Outcomes Research*, *10*(4), 351–353. doi:10.1586/erp.10.21

Squitieri, L., Bozic, K. J., & Pusic, A. L. (2017). The role of patient-reported outcome measures in value-based payment reform. *Value Health* Jun, *20*(6), 834–836. doi:10.1016/j.jval.2017.02.003, Epub 2017 Mar 22. PMID: 28577702; PMCID: PMC5735998.

Sul, D. (2024). Situating culturally specific assessment development within the disjuncture-response dialectic. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*. De Gruyter.

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., et al. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Quality of Life Research*, *27*(5), 1159–1170.

Türk, C., Neisius, A., Petrik, A., Seitz, C., Skolarikos, A., & Thomas, K. (2020). European Urological Association urolithiasis guidelines http://uroweb.org/guideline/urolithiasis/2020

U.S Dept of Health and Human Service Food and Drug administration.(2009). *Guidance for Industry: Patient reported outcome measures – Use in medical product development to support labelling claims*. Silver Spring: U.S. dept of health and Human Service: FDA.

US FDA and Scientific Advisory Committee of the Medical Outcomes Trust: Assessing health status and quality of life instruments: attributes and review criteria. Quality Research. (2002). The ability of the PROMs to improve decision-making relies on them accurately capturing the burden of disease or treatment.

Velikova, G., Booth, L., Smith, A. B., et al. (2004). Measuring quality of life in routine oncology practice improves communication and patient well-being: A randomized controlled trial. *Journal of Clinical Oncology*, *22*(4), 714–724. doi:10.1200/JCO.2004.06.078

Williams, K., Sansoni, J., Morris, D., Grootemaat, P., & Thompson, C. 2016. Patient-reported outcome measures: Literature review Australian commission on safety and quality in healthcare.

Wright, B. D., Mead, R. J., & Ludlow, L. H. (1980). KIDMAP: person-by-item interaction mapping (MESA Memorandum #29). Chicago:. MESA Press [http://www.rasch.org/memo29.pdf].

David Sul

15 Situating culturally specific assessment development within the disjuncture-response dialectic

Abstract: How can assessment be defined and operationalized to avoid measurement disjuncture, a misalignment wherein elements of an instrument development process from one worldview are applied in ways that negate and override another worldview? Culturally specific assessment is introduced to counter measurement disjuncture and to establish a critically derived form of assessment that is culturally responsive and relevant. The conceptual workspace of culturally specific assessment developers resides within a swirling environment of socio-historical factors that produces both disjunctures and cultural aspirations for new assessment possibilities, moving in uplifting directions. The theoretical framework of the disjuncture-response dialectic illuminates how settler colonialism and intellectual elimination contribute to measurement disjunctures and informs culturally specific assessment alternatives in support of intellectual amplification and indigenous sovereignty.

Keywords: culturally specific assessment, measurement disjuncture, Critical Theory, Indigenous knowledge, settler colonialism

15.1 Introduction

What role do large-scale assessment instruments contribute to preserving societal structures that limit the aspirations of Indigenous people? What role do sociohistorical factors play in the construction of large-scale assessment instruments administered to Indigenous people? How might the voices of Indigenous people be present and accounted for within large-scale assessment instruments? How can large-scale assessment research contribute to Indigenous peoples' aspirations for autonomy, self-determination, and liberation? This set of broad questions requires a broad perspective on the conceptual and physical spaces where assessment instrument development occurs. This chapter presents a measurement-based theoretical framework, the disjuncture-response dialectic (Sul, 2021), to establish the environment within which culturally specific assessment (Sul, 2019) development occurs. This framework situates the instrument development process within a swirling environment of sociohistorical factors that produces measurement disjuncture (Sul, 2019), cultural aspirations for new assessment possibilities, and uplifting directions for Indigenous communities (Sul, 2021).

David Sul, Office of the Provost, University of the Virgin Islands, Virgin Islands, USA

https://doi.org/10.1515/9783111036496-015

My work on the development of culturally specific assessments has been conducted in collaborative partnerships within Indigenous communities with Indigenous assessment developers who seek to center the Indigenous voices of those assessed. Such voices are often excluded from assessment theory and process development. This chapter honors my collaborators who helped the concepts presented here emerge.

15.2 Assessment alignment and misalignment

It is important to identify some key terms of this discussion. As such, this section describes assessment, three forms of assessment alignment, and a particular form of assessment misalignment.

15.2.1 What is assessment?

Sul (2021) compiled the core elements of various assessment definitions to describe assessment as:

> the representation of a domain of knowledge, skill, or affect through the use of procedures that allow for the translation of observations into assignments of value, permitting inferences about domain status for the purpose of making decisions. (Lynch, 2001; Popham, 2000; Thorndike & Thorndike-Christ, 2009)

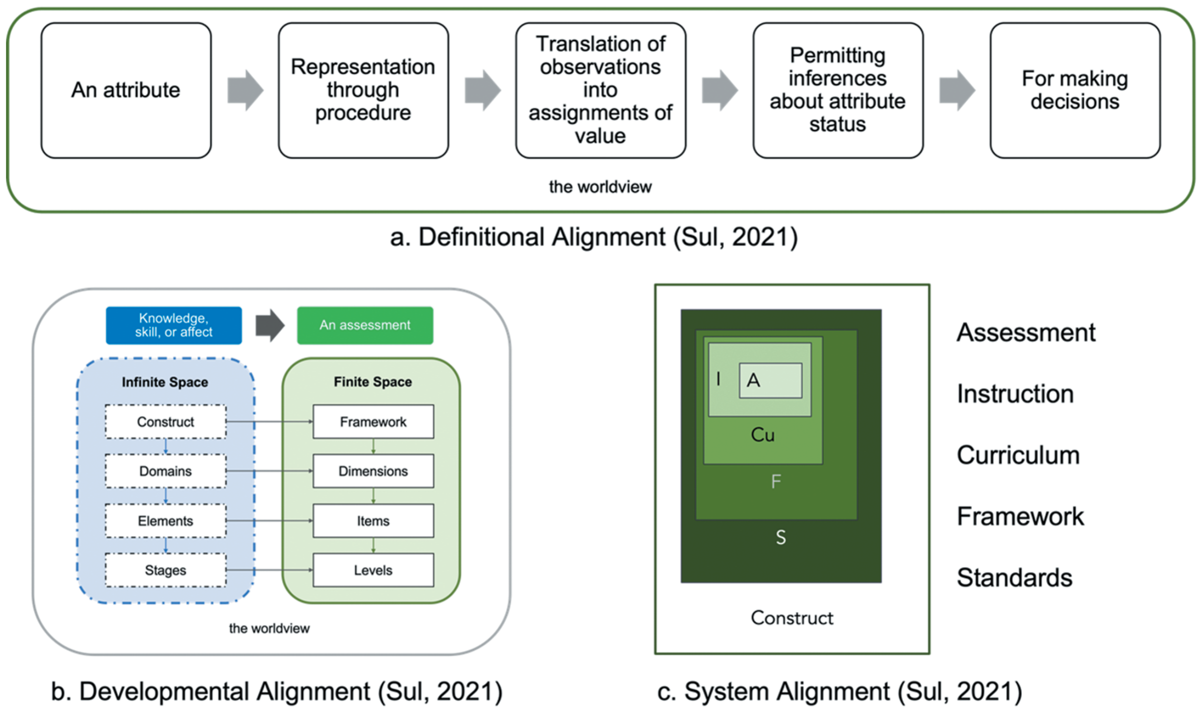
While the assignment of value may take either qualitative or quantitative form, it is important to note that under this definition, measurement is a form of assessment that relies on the quantification of the assigned value. The sense of quantification intended here is metrological, following Mari, et al. (2023), Pendrill (2019), and Wilson (2023). Though my purpose here is to spell out the sociohistorical aspects of culturally specific assessment, the measurement ideas and methods put to use in creating those assessments are in tune with efforts addressing the irreducible complexity involved in creating and sharing meaningful comparisons (Confrey et al., 2021; Fisher & Wilson, 2015; Lehrer, 2013; Mallinson, 2024).

15.2.2 Forms of assessment alignment

Assessment alignment is an often unstated objective of the instrument development process. It adds to the validity of the value assigned during the assessment process and can come in multiple forms. These include definitional alignment, developmental alignment, and system alignment (Sul, 2021). *Definitional alignment* (Figure 15.1a) is based on the above definition of assessment and expresses the need for the alignment of the five components of the definition. For example, standardized assessments frequently are developed from a dominant worldview that is often unnamed (Sul, 2021). Whether self-reported or through assessor ratings, the representation of a domain of knowledge, skill, or affect takes place within a cultural environment that necessarily assigns greater weight to some observations over others. Within academic environments within the United States, for example, the parameters for assessment development are frequently driven by the unstated dominant Western worldview. To maintain definitional alignment, instrument developers should "name their frame" and maintain alignment of the components of the assessment definition within that stated frame.

Developmental alignment (Figure 15.1b) describes alignment within and between the practical components of an assessment as operational projections of the conceptual components of an assessment. Under developmental alignment, the assessment construct is the knowledge, skill, or affect to be assessed and is operationalized as the assessment framework. Assessment domains represent the way the construct is subdivided, or not, and are operationalized as the assessment dimensions. The elements of the domains are the selected ways in which the domains are represented and are operationalized as the assessment items. Finally, stages are the developmental steps that guide progress within each of the domain elements and are operationalized as the assessment item rating levels. The conceptual elements of the assessment – the construct, domains, elements, and stages – each represent infinite ideas that are summarized into their respective practical finite ideas – the framework, dimensions, items, and levels – through the instrument development process. But where do these finite summaries of these infinite concepts reside? Within a worldview. To maintain developmental alignment, all conceptual and operational components should be aligned, both within and across the conceptual and practical frames, both of which should be contained within a named worldview.

System alignment (Figure 15.1c) is a form of assessment alignment that exists within, typically, educational systems that center content standards, content frameworks, forms of curricula, instruction, and, ultimately, assessment. Curricular content standards "represent ideas about what disciplinary content is most important for students to know and be able to do across years of schooling" (Wixson et al., 2003, p. 69). Content standards are also "ideological, reflecting values and beliefs regarding the nature of teaching and learning and, more generally, the purposes of education" (Wixson et al., 2003, p. 69). Under system alignment, curricular content standards are aligned with curricular content frameworks, forms of curricular instruction, and, ultimately, assessment (Sul, 2021). Alignment amongst each of these system components is highly regarded within educational systems. But whose standards drive these conversations about what should be taught, how it should be taught, and how it should be assessed?



a. Definitional Alignment (Sul, 2021)

b. Developmental Alignment (Sul, 2021)

c. System Alignment (Sul, 2021)

Figure 15.1: Forms of assessment alignment.

15.2.3 Measurement disjuncture as assessment misalignment

While various forms of assessment alignment are highly valued, the disjuncture-response dialectic (Sul, 2021) as a theoretical framework, came about in an attempt to understand a particular form of assessment misalignment that presents itself when assessments are designed and developed from within a dominant (e.g., Western) worldview and are applied within a marginalized (e.g., Indigenous) environment (Figure 15.2).

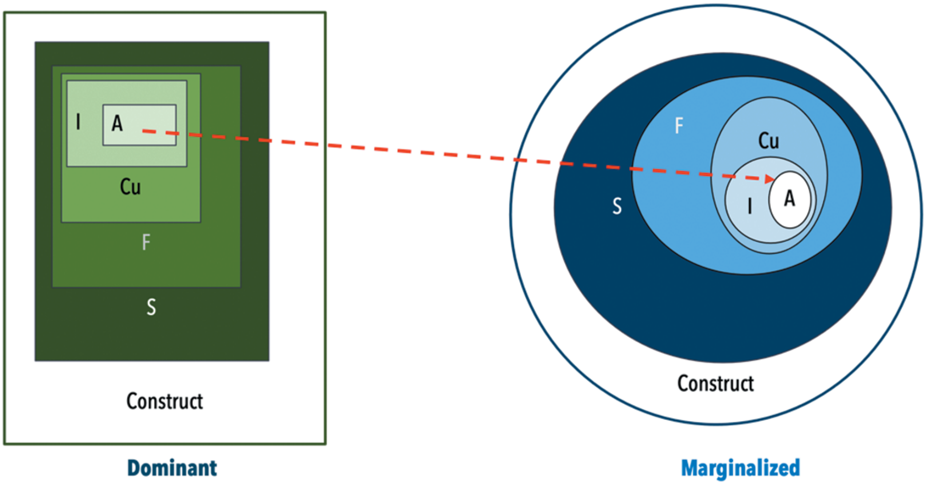


Figure 15.2: A particular form of assessment misalignment.

Throughout my assessment development work within Indigenous communities, it became necessary to understand what this form of assessment misalignment is, why it is a problem and what to do about it (Sul, 2019). Central to the description of the problem as it exists within Indigenous environments is that it relies on the ability of a dominant group to impose its assessments – and all that is required to design and develop them – onto another group that has been marginalized. As such, it was important to pursue a sociohistorical explanation for this form of misalignment. Sul (2019) examined the literature on settler colonialism (Wolfe, 1999, 2006), as this form of assessment misalignment within Indigenous environments seemed to be a reflection of colonialism. It is within that literature that two key terms were identified as potential descriptors of this special case of assessment misalignment.

Misalignment that is grounded in cultural and linguistic differences has been referred to as “disjuncture” (Appadurai, 1996; Meek, 2010; Wyman et al., 2010) and “discontinuity” (Bougie et al., 2003; Brown-Jeffy & Cooper, 2011; Edwards, 2006; Meek, 2007). Cultural discontinuity refers to the lack of cohesion between two or more cultures (Lovelace & Wheeler, 2006). Such cultural and linguistic disjunctures are often grounded in the conflicts of “beliefs, or feelings, about languages” that are the inevitable outcomes of the interaction of Indigenous, colonial, postcolonial, and professional academic perspectives (Kroskrity, 2009). Between these two descriptor options, Sul (2019) defined "the misalignment that occurs when elements of an instrument development process from one worldview are applied to the instrument development process of another worldview," as *measurement disjuncture* (Sul, 2019, p. 7).

15.2.4 Effects of measurement disjuncture

In a broad sense, measurement disjuncture affects the establishment of measurement validity, and, hence, the inferences made based on the information derived from assessment instruments (Sul, 2019). For example, how can one interpret the scores derived from an English-based computer adaptive assessment administered to students enrolled in a school that utilizes Hawaiian as the language of instruction? To understand its impact on measurement validity, it is important to review the formal definition of measurement validity. Measurement validity is the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, 2014). Key terms within this definition include "evidence," "theory," "interpretations," "scores," "uses," and "tests." The meanings of these terms within the very definition of measurement validity are grounded in and influenced by the worldview under which the instrument development occurs. When developing large-scale assessment instruments and competing definitions of the terms "evidence," "theory," "interpretations," "scores," "uses," and "tests" are present, it is far more likely for a dominant perspective on these terms to be used than for a perspective that emerges from a marginalized group.

Throughout the process of sharing measurement insights with groups across North America and beyond, the phrase "trying to fit a square peg into a round hole" was heard frequently to explain how Indigenous people feel about their interactions within the larger United States society, in general, and within educational environments, in particular. Disassembling the above phrase using imagery is instructive (Figure 15.3).

First, measurement disjuncture penalizes individuals from marginalized groups (e.g., Indigenous people) with limited exposure to the dominant culture and, hence, its influence on the assessment (Figure 15.3a). For Indigenous people, this can limit their capacity to attain credit for the dominant-worldview-based aspects of the assessment that are beyond their cultural reach. Additionally, because of measurement disjuncture, individuals from marginalized groups cannot receive credit for what they may know that exists outside of the dominant culture, upon which the assessment is based (Figure 15.3b). Depending on the assessment form, either of these two measurement disjuncture effects can lead to various misclassification errors. The result of these misclassification errors can be that Native American students are overrepresented (Type I error) in Special Education pro-

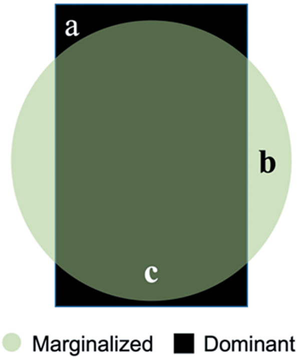


Figure 15.3: Effects of measurement disjuncture.

grams (Maureen E., 2016; Vining et al., 2017) or underrepresented (Type II error) in programs for gifted and talented students (Maker, 2020). Misclassification errors also result in the disproportional representation of Native American students receiving school discipline referrals (Brown, 2014; Whitford, 2017) and being "punished more harshly for lesser violations than their peers" (Brown, 2014; Gion et al., 2018). Gray, Brionez, Petros, and Gonzaga (2019) claimed that many psychological disorder assessments have been developed from within the Western worldview and culture, with the resulting effect being that others outside this worldview "may interpret questions differently, may have a different conceptualization of psychological wellness and illness as a whole, and may not share certain assumptions upon which such assessments implicitly or explicitly rely" (p. 534).

Finally, and most insidiously of all, as a result of measurement disjuncture, marginalized people must shift from their worldview to that of another, and essentially alter the complexion of their being in order to participate in the measurement activity (Figure 15.3c). This concept of masking one's true self within the assessment process is not new. It is reflective of the concept described as "double consciousness" by W.E.B Du Bois (1897, 2018) in his 1897 essay, "Strivings of the Negro People," and further explored in his 1903 book, *The Souls of the Black Folk*. The altering of one's sense of self to participate within dominant structures was expanded upon by Fanon (1952) in his seminal text, "Black Skin, White Masks." More recent researchers have referred to the "active denial of the present living existence of a culture and/or cultural identity as expressed through language, behaviors, norms, values, history, and assets" by educational structures as cultural identity silencing (Leigh-Osroosh & Hutchison, 2019, p. 2).

15.2.5 Prior attempts to describe measurement disjuncture

Attempts to describe the disjuncture within broader educational environments also are not new. Cultural discontinuity is defined conceptually as "a school-based behavioral process where the cultural value-based learning preferences and practices of many ethnic minority students – those typically originating from home or parental socialization activities – are discontinued at school" (Tyler et al., 2008, p. 281). The cultural discontinuity hypothesis, which originated in the ideas of anthropologists such as Dell Hymes (1974), posited that culture-based differences in the communication styles of minority students' home and the Anglo culture of the school lead to conflicts, misunderstandings, and, ultimately, failure for those students (Ledlow, 1992). Cultural discontinuity arises for students when their personal values clash with the ideals that shape their school system (Wiesner, 2006). Ladson-Billings (1995b) described the "discontinuity" problem as the gap between what students experience at home and what they experience at school with respect to their interactions of speech and language with teachers. Morris, Pae, Arrington, and Sevcik (2006) identified the most frequent roots of educational difficulties for Native American students as "the discontinuities between home and school in terms of language, culture, ideology, and educational expectations, which may be reinforced by incongruent instruction (pedagogy) and assessment methods or tools utilized in majority or mainstream schools" (p. 79). Since the 1990s, scholars have continued to discuss cultural discontinuity, variously terming it cultural mismatch (Ladson-Billings, 1995b), cultural incongruence (M. Foster et al., 2003), cultural misalignment (Tyler et al., 2006), cultural dissonance (Ladson-Billings, 1995b; Portes, 1999; Tillman, 2002), and cultural conflict (M. Foster et al., 2003).

15.3 Addressing measurement disjuncture through culturally specific assessment

This section focuses on the journey toward culturally specific assessment and describes prior attempts to address measurement disjuncture generally and within the assessment development process.

15.3.1 Prior attempts to address measurement disjuncture

Since the 1960s, a renaissance of the teaching of culture, language, and Indigenous knowledge has been occurring throughout Aotearoa (New Zealand), Hawai'i, Native American communities within the United States, and First Nations within Canada (Battiste, 2014; McCarty, 2003; van Meijl, 2006; Warschauer, 1998). Over this time, "Indigenous peoples and their allies have taken a stand and begun an indigenizing and decolonizing process" (P. Johnson, 2016, p. 45). These processes have included the retelling of cultural pasts and practices, advocacy for their own value systems, traditional forms of governance, and a return to ways of life that relate people to the cosmos, nature, and landscape.

In the 1970s, pressure on the United States federal government exerted by tribal nations and urban Indian communities within the United States focused on educational change and control. This led to “a number of important pieces of legislation and federal investigations related to American Indian education and, specifically, the role of tribal languages and cultures in schools serving Indigenous youth” (Brayboy & Castagno, 2009, p. 33). The Indian Education Act of 1972 was passed and included “opportunities and funding for creating tribal culture and language programs for schools and support for increasing the number of Native educators” (Brayboy & Castagno, 2009, p. 33). The challenges of educators trying to meet the needs of their Native American students resulted in additional federal legislation, Public Law 95–561 (P.L. 95–561) or the Indian Education Act of 1979, that included a call for a program of research and development of culturally specific assessments for use within Native American educational settings (Indian Education Act of 1979, 1982). Ten years after the passage of the Indian Education Act of 1979, Chavers and Locke (1989) wrote: “We do not know of any Native-normed test of any kind. This is an area that is obviously rich in development possibilities” (p. 19). In 1995, Estrin and Nelson-Barber (1995) wrote “there is no repertoire of standardized tests in Native languages or that draw on Native cultural content and learning processes” (p. 5).

Prior attempts were made to address the “discontinuity” problem (Ladson-Billings, 1995). Au and Jordan (1981) described as “culturally appropriate” the incorporation of “talk story” into a program of reading instruction for Native Hawaiian students that improved upon expected scores on standardized reading tests. Mohatt, Erickson, Trueba, and Guthrie (1981) used the term “culturally congruent” to describe teachers’ use of interaction patterns that simulated Native American students’ home cultural patterns to produce improved academic performance. Jordan (1985) defined educational practices as “culturally compatible” when the culture of students is used as a guide in choosing aspects of the educational program to maximize academically desired behaviors and minimize undesired behaviors. Researchers, beginning in the 1980s, used the term “culturally responsive education” to describe the language interactions of teachers with linguistically diverse and Native American students (Cazden & Leggett, 1981; Erickson & Mohatt, 1982). Erickson and Mohatt (1982) suggested their notion of culturally responsive teaching could be seen as a beginning step for bridging the gap between home and school. Ladson-Billings (1995b) claimed the term culturally responsive represented a more expansive, dynamic, and synergistic relationship between the culture of the school and that of the home and greater community.

A growing international Indigenous rights movement led to the passage of Article 14.1 of the 2007 United Nations Declaration on the Rights of Indigenous Peoples. It asserted “Indigenous peoples have the right to establish and control their educational systems and institutions, providing education in their own languages, in a manner appropriate to their cultural methods of teaching and learning” (United Nations Declaration on the Rights of Indigenous Peoples, 2007, p. 5). Since then, Indigenous communities have reframed their educational settings (Ragoonaden & Mueller, 2017) to align with their cultural worldviews and within these settings resides the practice of formal assessment. According to Brayboy and Castagno (2009), “two models dominate conversations and approaches to Indian education in the USA: the assimilative model and the culturally responsive model" (p. 31). In addition to developing teaching materials and resources, Indigenous scholars such as Sʔímlaʔxʷ Michele K. Johnson now call on Indigenous educators to "create their own methods of assessing student achievement and fluency" (2017, p. 23).

15.3.2 Responding to measurement disjuncture in assessment development

The lack of representation of Indigenous perspectives within assessment development processes has been met by a range of techniques. One method of developing assessments is to begin with one that already has been validated for one setting and then modify it for use in another (Borgia, 2009). Given the challenge of assessing Indigenous knowledge domains using existing assessments, some have focused their efforts on the development of entirely new assessments grounded in the perspectives of Indigenous people (Dench et al., 2011). A typical psychometric response to assessment misfit would be to continue to use the assessment or a modified version of it and to examine such issues as internal consistency, item bias, and differential item functioning for Native American students. McGroarty, Beck, and Butler (1995) wrote that such responses have focused on the "technical and statistical properties of language assessments and excluded consideration of wider educational and human consequences" (p. 323). Others have indicated that when working within Indigenous settings, "it is sometimes not possible to do a full evaluation of psychometric properties such as reliability, validity, and sensitivity" (Dench et al., 2011, p. 171) due to the relatively small population sizes.

15.3.3 Arriving at culturally specific assessment

There have been multiple attempts to address educational disjuncture issues written about in the research literature. Ladson-Billings (1995b) conducted a field-altering qualitative study on the teaching methods of teachers who demonstrated consistent academic success with African American students. Ladson-Billings (1995b), grounded in Black feminist thought, introduced the theory of "culturally relevant pedagogy" to emphasize the significance of teaching to and through the cultural strengths of ethnically diverse students. Ladson-Billings (1995b) and Jordan (1985) argued for the use of culturally relevant pedagogy to engage actively and motivate students from ethnically diverse backgrounds to improve their academic achievement. Ladson-Billings (1995b) established three criteria for a culturally relevant pedagogy that could be used to address the "discontinuity" problem: (a) an ability to develop students academically, (b) a willingness to nurture and support cultural competence to help students to maintain their cultural integrity while succeeding academically, and (c) the development of a sociopolitical or critical consciousness. In a culturally relevant classroom, a child's culture is not only acknowledged but also seen as a source of strength that can be utilized to attain academic success.

Sociopolitical consciousness has been described as an individual's ability to analyze critically the political, economic, and social forces shaping society and one's status in it (Seider et al., 2018). For her last definitional criterion, Ladson-Billings (1995b) borrowed from Freire (2017) and acknowledged that students must develop a broader sociopolitical consciousness and the skills to critique the cultural norms, values, mores, and institutions that produce and maintain social inequities. The development of a sociopolitical or critical consciousness within students allows them to acknowledge and act on historical circumstances that affect their current reality (Freire, 2017; Ladson-Billings, 1995b). As such, when culturally relevant pedagogy is conducted within North America, the aftereffects of colonialism and slavery must be taken into consideration in order to develop sociopolitical or critical consciousness within students. Critical consciousness is defined here as an awareness of and desire to act against societal inequities that disadvantage learners. Critical consciousness researchers acknowledge the key role that education can play in dismantling societal inequalities. This requires the deconstruction of the assessment development processes and the identification of sources of potential discontinuities that arise between conflicting epistemologies, constructs, representations of the construct, notions of what is considered measurable, and methods of measurement.

Researchers in the field of program evaluation began to utilize the term "responsive evaluation" in the early 1970s, in reference to a focus on issues of practical importance to program managers and developers (Stake, 2011). Stake (1973) sought to remove the emphasis on static program objectives developed by those furthest from the delivery of program services and stressed the importance of being responsive to situational realities in the management of programs and to the reactions, concerns, and issues of participants. This represented a dramatic departure from the emphasis on the use of evaluation plans that relied on preconceived notions of program expectations. Stake (1973) believed that the ultimate test of the validity of an evaluation is the extent to which it increases the audience's understanding of the program. Stake's (1973) work led to the stream of responsive evaluation research and practices that exist today.

Drawing upon the lineage of research in responsive evaluation and culturally relevant pedagogy, Hood (1998) argued that student learning is assessed more effectively using assessment approaches that are culturally responsive. Combining the ideas of Ladson-Billings (1995) and Stake (1973), Hood (1998) promoted the development of "culturally responsive" performance-based assessments as a means of achieving equity for students of color. Hood (1998) noted that there were to be challenges and difficulties in the development of both performance tasks and scoring criteria that would be "responsive to cultural differences and adequately assess the content-related skills that are the focus of the assessment" (p. 192). In the case where a culturally responsive assessment minimizes measurement disjuncture by allowing learners to be present as their full and complete selves within the assessment activity and to receive maximum credit for the things they know that exist outside of the dominant culture upon which the assessment is based, measurement disjuncture still penalizes learners with limited exposure to the dominant culture and, hence, its influence on the assessment.

In examining these various responses to measurement disjuncture, I focused initially on the culturally responsive assessment approach proposed by Dr. Stafford Hood. Hood (1998) established culturally responsive assessment, not as a means to address measurement disjuncture per se, but it was important to rely on an existing theoretical assessment model for how to respond to measurement disjuncture. In electronic mail communication with Dr. Hood about my work in Hawaiʻi, he noted that what I was working on there was not culturally responsive assessment but culturally specific assessment (S. Hood, personal communication, May 13, 2018). There really was only one additional characteristic that was needed to supplement Hood's definition of culturally responsive assessment (Hood, 1998) and that was the identification and naming of the worldview within which the culturally specific assessment development work takes place (Sul, 2019). The result was the formal definition of culturally specific assessment as assessment that (a) supports the (academic) development of individuals, (b) is inclusive of a willingness to nurture and support cultural competence, (c) aims to support the development of a sociopolitical or critical consciousness within students, (d) is focused on constructs and measures of importance to educational practitioners and other key stakeholders, and (e) functions within a system of knowledge that exists within a named worldview (Sul, 2019).

15.3.4 Grounding culturally specific assessment

An "emic" approach (Hui & Triandis, 1985) to research, as opposed to an etic approach, refers to research that studies phenomena that exist within one culture and does not involve a focus on other cultures. These two terms are derived from linguistics where "phonetics refers to the study of general aspects of vocal sounds and their production and phonemics studies the sounds used in a particular language" (Eckensberger, 2015, pp. 111–112). The etic research approach refers to research when it is conducted "across many cultures, when the structure is created, and when the criteria for analysis are considered absolute or universal" (Eckensberger, 2015, p. 112). The main aim of the emic approach, located at one end of the "abstraction universality-cultural specificity continuum" (Hui & Triandis, 1985, p. 132), is to focus on individual differences in attributes that are characteristic of a cultural context (Burtăverde et al., 2018). This emic approach has been applied in a variety of disciplines such as cancer prevention (Garcia et al., 2017), student behavior (Hitchcock et al., 2005), early-childhood education (Kinzel, 2015), and mental health (O’Brien et al., 2007; Telander, 2012; The Getting it Right Collaborative Group et al., 2019; Thompkins et al., 2020; Walls et al., 2016; Whitfield, 2017).

Nastasi (2000) wrote that educational psychological services that are culturally specific “embody an individual’s real-life experiences within a given cultural context . . . and his or her understanding of those experiences” (p. 547). A core aspect of the culturally specific approach is the cultural lens through which the culturally specific assessments are developed. I want to be clear that this lens could be singular or multifaceted. If it is the latter, this aligns quite well with the concept of intersectionality (Crenshaw, 1991). Finally, defined broadly, the cultural worldview may reflect a linguistic culture or a thematic culture. In fact, culturally specific assessment can be applicable anywhere groups of individuals congregate together and establish a culture that reflects their shared values and interests.

15.3.5 Situating culturally specific assessment within a critical theoretic taxonomy

Identification of the five components of the definition of culturally specific assessment allows for the construction of the critical theory (Horkheimer, 2018) taxonomy that houses them. Working backward from culturally specific assessment to culturally responsive assessment (Hood, 1998) led directly to the work of Gloria Ladson-Billings (1995), who relied on the Freirean notion of developing a critical consciousness (Freire, 1970), within her definition of culturally relevant teaching. From Freire, it is easy to link back to the work of the critical theorists of the 1930s, 1920s, and earlier. The exercise of establishing the critical taxonomy wherein culturally specific assessment resides allows for its visual placement within an organizational chart of critical theories (Figure 15.4). Within this chart, culturally specific assessment and the disjuncture-response dialectic (Sul, 2021) are situated with respect to other related critical theories such as Critical Race Theory (Bell, 1995) and QuantCrit (Gillborn et al., 2018).

Seeking a more concise way to think about and represent where culturally specific assessment resides, I went back to a very familiar workspace for me, the Venn diagram (Venn, 1971), and added critical theory, juxtaposed against both assessment and measurement as concentric concepts. Critical theorists are often at odds even within their own disciplines and to reflect this, I chose to construct a *Critical Venn Diagram* – one that juxtaposes critical theory against disciplines of interest – using an orthogonally overlapping oval instead of another circle, to draw a distinction between the typical side-by-side pairing of circles found in most Venn diagrams. The critical Venn diagram (Figure 15.5) allowed for the crossing of critical theory against both measurement and assessment and to establish, visually, the critical theory-infused environment wherein critical assessment (Figure 15.5a), and critical measurement (Figure 15.5b) reside. Culturally responsive assessment (Figure 15.5c) resides within critical pedagogy (Figure 15.5a) and within culturally responsive assessment lies culturally specific assessment (Figure 15.5d). An impor-

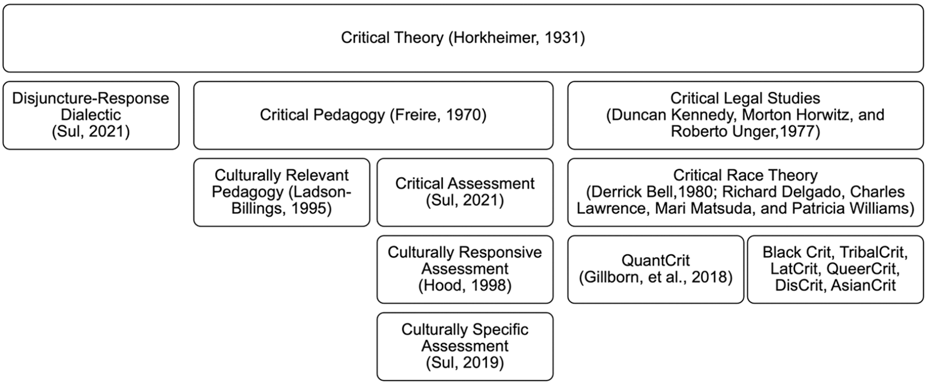

Critical Theory (Horkheimer, 1931)
Disjuncture-Response Dialectic (Sul, 2021)
Critical Pedagogy (Freire, 1970)
Critical Legal Studies (Duncan Kennedy, Morton Horwitz, and Roberto Unger,1977)
Culturally Relevant Pedagogy (Ladson-Billings, 1995)
Critical Assessment (Sul, 2021)
Critical Race Theory (Derrick Bell,1980; Richard Delgado, Charles Lawrence, Mari Matsuda, and Patricia Williams)
Culturally Responsive Assessment (Hood, 1998)
QuantCrit (Gillborn, et al., 2018)
Black Crit, TribalCrit, LatCrit, QueerCrit, DisCrit, AsianCrit
Culturally Specific Assessment (Sul, 2019)

Figure 15.4: Taxonomy of critical theories.

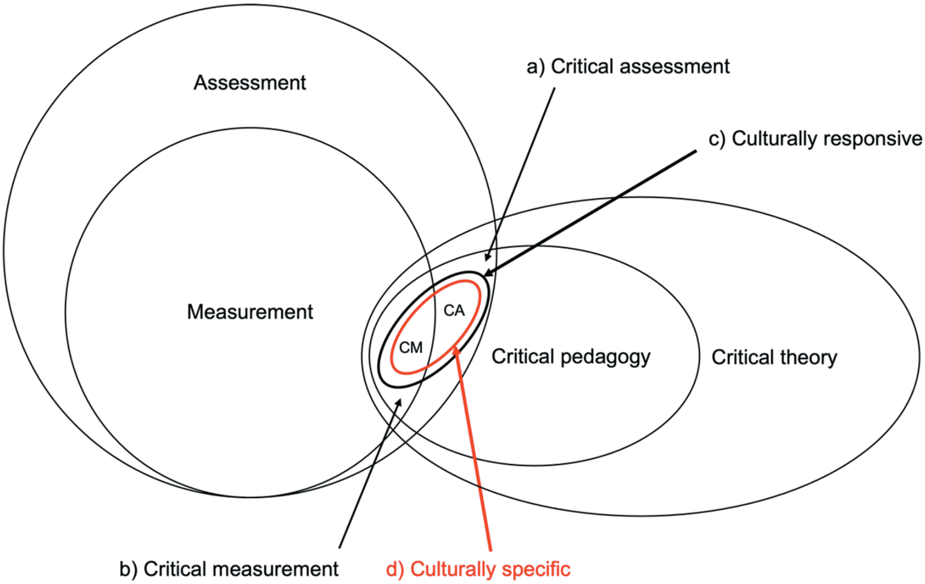


Figure 15.5: Situating culturally responsive and culturally specific assessment.

tant aspect of the culturally specific approach is the one or more lenses through which the culturally specific assessments are developed. This lens is represented by the outline coloring of the culturally specific assessment space.

With the establishment of the taxonomy of a host of critical theories and the visual representation of the location of culturally specific assessment, the core question related to the work environment of culturally specific assessment developers remains: What are the swirling forces that comprise the environment within which culturally specific assessment developers do their work? That theoretical workspace is framed by the disjuncture-response dialectic (Sul, 2021), which provides a description of those forces at play when culturally specific assessment developers do their work. In fact, these forces go beyond the construction of culturally specific assessments and can be applied to a number of research concepts, beyond assessment. This led to the establishment of the generalized disjuncture-response dialectic (Sul, 2021), a generalized space where researchers who are attempting to develop culturally specific responses to various disjunctures do their work.

15.4 The disjuncture-response dialectic

Nineteenth-century German philosopher Georg Wilhelm Friedrich Hegel (2010) wrote, “contradiction is the root of all movement and vitality; it is only insofar as something has a contradiction within it that it moves, has an urge and activity” (p. 439). The act of exposing the contradiction serves as the impetus for the emergence of the next iteration of the concept, idea, or framework. This concept is actualized when the imposition of non-Indigenous forms of assessment onto Indigenous people leads to a disjuncture (Appadurai, 1996; Meek, 2010; Wyman et al., 2010) and a corresponding response that is multilayered and affects all aspects of the work of Indigenous assessment developers.

The disjuncture-response dialectic (Sul, 2021) is a theoretical framework that situates the work of culturally specific assessment developers within a swirling environment of sociohistorical factors, cultural aspirations, and uplifting directions. The instruments developed within this environment contain an acknowledgement of the historical legacy of slavery, institutional racism, settler colonialism (Wolfe, 1999, 2006), intellectual elimination (Sul, 2021), and their impact on measurement disjuncture (Sul, 2019). Simultaneously, the work of culturally specific assessment developers serves as a political act of intellectual amplification and liberation that challenges intellectual elimination (Sul, 2021).

The generalized disjuncture-response dialectic explains the multitiered environment wherein all culturally specific researchers function (Sul, 2021). The layers of the disjuncture-response dialectic (Figure 15.6) and the elements of it are described in the sections below.

15.4.1 Settler colonialism as sociohistorical factor

The deep relationship between Indigenous people and their specific forms of assessment was disrupted by European contact. Within Indigenous communities, that disruption must be acknowledged and addressed before the practice of assessment development for use with Indigenous people can begin.

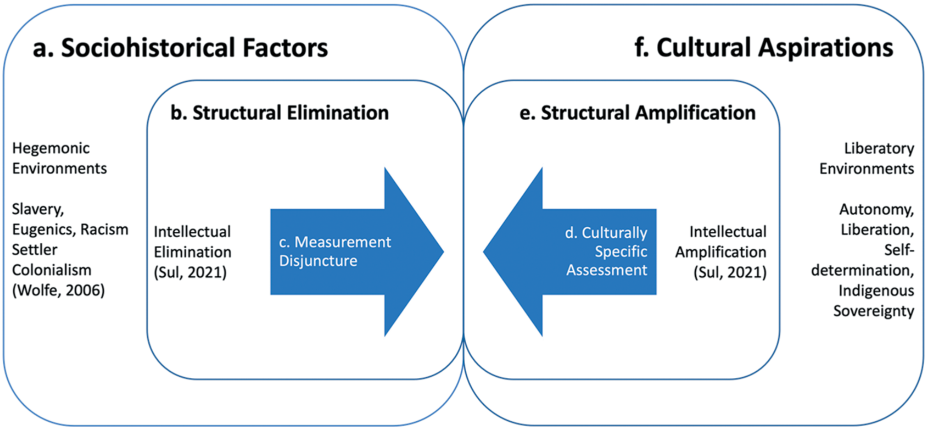

a. Sociohistorical Factors
b. Structural Elimination
Hegemonic Environments
Slavery, Eugenics, Racism Settler Colonialism (Wolfe, 2006)
Intellectual Elimination (Sul, 2021)
c. Measurement Disjuncture
f. Cultural Aspirations
e. Structural Amplification
d. Culturally Specific Assessment
Intellectual Amplification (Sul, 2021)
Liberatory Environments
Autonomy, Liberation, Self-determination, Indigenous Sovereignty

Figure 15.6: The disjuncture-response dialectic.

Indigenous people continue to experience the effects of colonialism and the "decimation of the indigenous population, primarily through waves of disease, annihilation, military and colonialist expansionist policies" (Brave Heart & DeBruyn, 1998, p. 62). Indigenous people have been subjected to historical and contemporary complexities such as "genocide, territorial usurpation, forced relocation, and transformations of Native economic, cultural and social systems brought on by contact with Whites" (McCarty, 2003, p. 148). In Hawai'i, there were an estimated 800,000 Hawaiians prior to the arrival of Captain Cook in 1778 and within 100 years, venereal diseases, tuberculosis, and influenza decimated nearly 95% of the Native Hawaiian population (Warner, 1999). In North America, European colonization "forced North American tribes from their ancestral homelands, destroyed their communities (culturally and literally), and forced assimilation to a European way of life that is now considered mainstream North American culture" (Bowman et al., 2015, p. 337). Indigenous people continue to be harmed by historical trauma, the chronic trauma and "unresolved grief of a people due to systemic loss" (Shea et al., 2019, p. 554) that affects both survivors and subsequent generations (Brave Heart & DeBruyn, 1998; Grayshield et al., 2015; Morgan & Freeman, 2009).

Colonialism, according to Yellow Bird (1999), takes place when an alien people invade the territory inhabited by people of a different race and culture, and establish political, social, spiritual, intellectual, and economic domination over that territory. It includes the appropriation of both territory and resources by the colonizer and loss of sovereignty by the colonized (Yellow Bird, 1999). Patrick Wolfe (2006) defined settler colonialism as inherently eliminatory but not invariably genocidal. Wolfe (2006) described the logic of elimination as the summary liquidation of Indigenous people and their societies. As with genocide, settler colonialism first strives for "the dissolution of native societies" and, then, the construction of "a new colonial society on the expropriated land base" (p. 388). According to Wolfe (2006), the primary motive for elimination "is not race (or religion, ethnicity, grade of civilization, etc.), but access to territory" (p. 388).

Sul (2021) interviewed Indigenous assessment developers and gathered their reflections to understand how they acknowledge the presence of settler colonialism (Wolfe, 2006) within their assessment development work.

15.4.2 Intellectual elimination as structural elimination

Applying a modified form of Wolfe's concept of the logic of elimination (Wolfe, 2006), Sul (2021) constructed the logic of *intellectual* elimination as also being inherently eliminatory. The logic of *intellectual* elimination refers to the summary dissolution of native societies' *knowledge* and then for the construction of a new colonial *knowledge* within the expropriated *minds*. As with the logic of elimination, the primary motive for intellectual elimination is not race (or religion, ethnicity, grade of civilization, etc.), but access to territory (Sul, 2021). Sul (2021) further posited that assessment developers who practice Western forms of assessment development within Indigenous communities are participants in this intellectual elimination. Sul (2021) presented multiple cases to demonstrate the relative ease with which assessment developers and researchers introduce intellectual colonialism through their practices and methods. The consequences of their actions are incalculable. Three such cases are presented here.

In use throughout Canada, parenting capacity assessments (PCA) are used by child protection workers to make determinations about the fitness of parents to care for their children (Choate & McKenzie, 2015). In noting the role that neglect investigations play in the overrepresentation of Indigenous children in child welfare, Caldwell and Sinha (2020) called for a "framework for reform of current approaches to assessing and addressing cases involving concerns about neglect" (p. 483). When making important decisions about child protection, Muir and Bohr (2014) argue that "the cultural, social and historical realms of Aboriginal communities" must be considered in the assessment of Aboriginal children, "especially in the context of child protection, as identifiable differences may exist between the parenting norms in Aboriginal communities and those of mainstream groups" (p. 76). Nevertheless, PCAs in use throughout Canada are a part of larger decision-making processes that "have been constructed using Euro-North America understandings of parenting, focusing on the nuclear family" (Choate & McKenzie, 2015, p. 32).

The Lakota Women and Cervical Cancer Survey (Bowker, 2017; Bowker et al., 2020) was developed to conceptualize the knowledge, beliefs, and behaviors of Lakota women with respect to the Human Papillomavirus (HPV) and cervical cancer. Lakota women have their own distinct worldview and beliefs about health and yet the survey included slight modifications to a previously developed instrument constructed for use with Appalachian women (Vance & Keele, 2013).

Mental health screenings and assessments that are not responsive to the needs of Latinx immigrants are used frequently for evaluations of clinical programs that serve them (Alegría et al., 2019; Cardemil et al., 2010; Farina & Mancini, 2017; Kaltman et al., 2016; Kataoka et al., 2003; Santiago et al., 2015). When Latinx immigrants present for trauma care within these mental health programs, they are often assessed with culturally encapsulated (McCubbin & Bennett, 2008) instruments that fail to capture: (a) Latinx cultural experiences, values, and knowledge, (b) the specific forms of pre-migration, during migration, and postmigration traumas they may encounter, and (c) how colonization, enslavement, racism, and other oppressive forces shape their experiences (Sul & Domínguez, 2021). In a review of the six evaluation studies cited above, 23 unique mental health assessment instruments were utilized: Patient Health Questionnaire (PHQ-9); Generalized Anxiety Disorder Scale (GAD-7); Posttraumatic Stress Disorder (PTSD) Checklist (PCL-5); Hopkins Symptom Checklist (HSCL-20); Beck Depression Inventory (BDI); Farina, A. S. J., & Mancini, M. (2017); UCLA PTSD Index for DSM-IV; Screen for Child Anxiety Related Disorders (SCARED); National Institute of Mental Health Center for Epidemiological Studies Depression Scale for Children (CESDC); Pediatric Symptom Checklist (PSC); The Physiological Hyperarousal Checklist; The Emotion Regulation Checklist (ERC); Stressful Life Events Screening Questionnaire (SLESQ); PTSD Checklist; Medical Outcomes Study Social Support Survey (MOS-SSS); Life Events Scale Child; PTSD Symptom Scale (CPSS), the child version of the Posttraumatic Diagnostic Scale for Adults (PDSA); Children's Depression Inventory (CDI); Modified version of the Attitudes Toward Mental Health Treatment Scale (ATMHT); Parental Involvement in School measure; Responses to Stress Questionnaire (RSQ); Family Crisis Oriented Personal Evaluation Scales (FCOPES); Familism Scale (Gil, Wagner, & Vega, 2000); and the Child Report of Parenting Behavior Inventory (CRPBI-Parent version).

Although some of these evaluation researchers attempted to be responsive to cultural and linguistic needs of the immigrant participants during the assessment process, this responsiveness began and ended with a strict Spanish-language translation of the instrument.

Sul (2021) interviewed Indigenous assessment developers and gathered their reflections to understand how they acknowledge the presence of intellectual elimination within their assessment development work.

15.4.3 Intellectual amplification as structural amplification

The term "amplification" came to me from the realm of mathematics and is a term familiar to people who listen to music, play an electric instrument, or have attended an outdoor concert where amplifiers are used. Amplifiers take a smaller sound and make it larger. The term "intellectual amplification" (Sul, 2021), thus, is intended to convey the various ways in which Indigenous voices can be heard, in response to intellectual elimination. Intellectual amplification begins with the acknowledgement that Indigenous culture, language and knowledge systems exist. It also includes revitalization, sustenance, maintenance, development, and the promotion of culture, language, and knowledge systems. According to Battiste (2005):

> Whether or not it has been acknowledged by the Eurocentric mainstream, Indigenous knowledge has always existed. The recognition and intellectual activation of Indigenous knowledge today is an act of empowerment by Indigenous people. The task for Indigenous academics has been to affirm and activate the holistic paradigm of Indigenous knowledge to reveal the wealth and richness of Indigenous languages, worldviews, teachings, and experiences, all of which have been systematically excluded from contemporary educational institutions and from Eurocentric knowledge systems. (p. 1)

McCarty and Lee (2014) wrote that culturally sustaining/revitalizing pedagogy (CSRP) addresses “sociohistorical and contemporary contexts of Native American schooling,” “attends directly to asymmetrical power relations and the goal of transforming legacies of colonization,” “recognizes the need to reclaim and revitalize what has been disrupted and displaced by colonization,” and “recognizes the need for community-based accountability” (p. 103). Culturally specific assessment aligns with a CSRP approach.

Sul (2021) interviewed Indigenous assessment developers and gathered their reflections on the concept of intellectual amplification and how their work contributed to it.

15.4.4 Indigenous sovereignty as cultural aspiration

While intellectual amplification can come in many forms from a variety of cultural worldviews, when gathered across Indigenous groups, amplification of Indigenous knowledge forms but one strategy within broader political movements that seek the full expression of the right to Indigenous sovereignty. Sovereignty is the right of a people to self-government, self-determination, and self-education, which includes the right to linguistic and cultural expression according to local languages and norms (Lomawaima & McCarty, 2002). According to Lomawaima (2000), the sovereignty held by Native American tribes has inherently existed prior to the establishment of the United States and is the “bedrock upon which any and every discussion of Indian reality today must be built” (p. 3).

The drive toward Indigenous sovereignty is where the work of Indigenous culturally specific assessment developers resides. As such, Indigenous culturally specific assessment developers are political actors and their assessment development practices, offered in response to measurement disjuncture, serve as political acts of intellectual amplification and Indigenous sovereignty that challenge intellectual elimination, and, ultimately, stand against forces of settler colonialism.

Sul (2021) interviewed Indigenous assessment developers and gathered their reflections to understand how these developers believe their work contributes to the grander goal of Indigenous sovereignty.

15.5 Discussion

Culturally specific assessment development commences with the identification and articulation of external factors and aspirations influencing the work of the culturally specific assessment developers. The resulting culturally specific instrument development process acknowledges the historical legacy of slavery, institutional racism, settler colonialism (Wolfe, 1999, 2006), intellectual elimination (Sul, 2021), and their impact on measurement disjuncture (Sul, 2019). Culturally specific assessment is a form of intellectual amplification and elevates Indigenous cultural aspirations such as liberation and freedom (Sul, 2021).

Through the design and development of culturally specific assessment instruments, measurement researchers and assessment instrument developers have great power to influence change within Indigenous communities. By interrogating the historical practice of measuring population-order phenomena within Indigenous communities using instruments that have been developed using dominant frames of reference, measurement researchers and assessment instrument developers can center a diversity of Indigenous worldviews, previously excluded from measurement research. To do so, measurement researchers can decouple instrument development from reliance upon dominant Western perspectives and intentionally create spaces for Indigenous perspectives regarding what should be measured and how it should be measured. The result can be the design and development of assessment instruments that are grounded in Indigenous ontologies, integrating the voices of those assessed throughout the instrument development process, and contribute to the liberatory aspirations of Indigenous people.

References

Alegría, M., Falgas-Bague, I., Collazos, F., Carmona Camacho, R., Lapatin Markle, S., Wang, Y., Baca-García, E., Lê Cook, B., Chavez, L. M., Fortuna, L., Herrera, L., Qureshi, A., Ramos, Z., González, C., Aroca, P., Albarracín García, L., Cellerino, L., Villar, A., Ali, N., & Shrout, P. E. (2019). Evaluation of the integrated intervention for dual problems and early action among Latino immigrants with co-occurring mental health and substance misuse symptoms: A randomized clinical trial. *JAMA Network Open*, *2*(1), e186927. https://doi.org/10.1001/jamanetworkopen.2018.6927

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Appadurai, A. (1996). *Modernity at large: Cultural dimensions of globalization*. University of Minnesota Press.

Battiste, M. (2005). Indigenous knowledge: Foundations for first nations. *WINHEC: International Journal of Indigenous Education Scholarship*, *1*, 1–17.

Battiste, M. (2014). Ambidextrous epistemologies: Indigenous knowledge within the Indigenous Renaissance. In S. Kamboureli & C. Verduyn (Eds.). *Critical collaborations: Indigeneity, diaspora, and ecology in Canadian literary studies* (pp. 83–98). Wilfrid Laurier University Press.

Bell, D. A. (1995). Who's afraid of critical race theory. *University of Illinois Law Review*, *1995*(4), 893–910.

Borgia, M. (2009). Modifying assessment tools for Ganöhsesge:kha: Hë:nödeyë:stha a Seneca culture-language school. In J. A. Reyhner & L. Lockard (Eds.). *Indigenous Language Revitalization: Encouragement, guidance & lessons learned* (pp. 191–210). Northern Arizona University.

Bougie, É., Wright, S. C., & Taylor, D. M. (2003). Early heritage-language education and the abrupt shift to a dominant-language classroom: Impact on the personal and collective esteem of Inuit children in Arctic Québec. *International Journal of Bilingual Education and Bilingualism*, *6*(5), 349–373.

Bowker, D. (2017). *Knowledge and beliefs regarding HPV and cervical cancer among Lakota women living on the Pine Ridge Indian Reservation and cultural practices most predictive of cervical cancer preventative measures* [Doctoral dissertation].

Bowker, D., Gee, J., & Huttlinger, K. (2020). Development of a culturally valid instrument examining HPV knowledge and beliefs of Lakota women on the pine ridge reservation. *Journal of Transcultural Nursing*, 104365962094780. https://doi.org/10.1177/1043659620947809

Bowman, N. R., Francis, C. D., & Tyndall, M. (2015). Culturally responsive Indigenous evaluation. In S. Hood, R. K. Hopson, & H. Frierson (Eds.). *Continuing the journey to reposition culture and cultural context in evaluation theory and practice* (pp. 335–359). Information Age Publishing.

Brave Heart, M. Y. H., & DeBruyn, L. M. (1998). The American Indian holocaust: Healing historical unresolved grief. *American Indian and Alaska Native Mental Health Research*, *8*(2), 60–82. https://doi.org/10.5820/aian.0802.1998.60

Brayboy, B. M. J., & Castagno, A. E. (2009). Self-determination through self-education: Culturally responsive schooling for Indigenous students in the USA. *Teaching Education*, *20*(1), 31–53. https://doi.org/10.1080/10476210802681709

Brown, C. A. (2014). Discipline disproportionality among American Indian students: Expanding the discourse. *Journal of American Indian Education*, *53*(2), 29–47.

Brown-Jeffy, S., & Cooper, J. E. (2011). Toward a conceptual framework of culturally relevant pedagogy: An overview of the conceptual and theoretical literature. *Teacher Education Quarterly*, *38*(1), 65–84.

Burtăverde, V., De Raad, B., & Zanfirescu, A.-Ş. (2018). An emic-etic approach to personality assessment in predicting social adaptation, risky social behaviors, status striving and social affirmation. *Journal of Research in Personality*, *76*, 113–123. https://doi.org/10.1016/j.jrp.2018.08.003

Caldwell, J., & Sinha, V. (2020). (Re) conceptualizing neglect: Considering the overrepresentation of Indigenous children in child welfare systems in Canada. *Child Indicators Research*, *13*(2), 481–512. https://doi.org/10.1007/s12187-019-09676-w

Cardemil, E. V., Kim, S., Davidson, T., Sarmiento, I. A., Ishikawa, R. Z., Sanchez, M., & Torres, S. (2010). Developing a culturally appropriate depression prevention program: Opportunities and challenges. *Cognitive and Behavioral Practice*, *17*(2), 188–197. https://doi.org/10.1016/j.cbpra.2010.01.005

Cazden, C., & Leggett, E. (1981). *Culturally responsive education: Recommendations for achieving Lau remedies II*. U.S. Department of Health, Education & Welfare, National Institute of Education.

Chavers, D., & Locke, P. (1989). *The effects of testing on Native Americans* (p. 49) [Research/Technical]. Native American Scholarship Fund, Inc.

Choate, P. W., & McKenzie, A. (2015). Psychometrics in parenting capacity assessments: A problem for aboriginal parents. *First Peoples Child & Family Review*, *10*(2), 31–43.

Confrey, J., Shah, M., & Toutkoushian, E. (2021). Validation of a learning trajectory-based diagnostic mathematics assessment system as a trading zone. *Frontiers in Education*, *6*, 654353. https://doi.org/10.3389/feduc.2021.654353

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*(6), 1241–1299. https://doi.org/10.2307/1229039

Dench, C., Cleave, P. L., Tagak, J., & Beddard, J. (2011). The development of an Inuktitut and English language screening tool in Nunavut. *Canadian Journal of Speech-Language Pathology and Audiology*, *35*(2), 168–177.

Du Bois, W. E. B. (1897). Strivings of the Negro people. *Strivings of the Negro People*. *80*, 194–198. Ignacio: USF Libraries Catalog.

Du Bois, W. E. B. (2018). *The souls of black folk*. Myers Education Press, LLC; Ignacio: USF Libraries Catalog. https://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=cat00548a&AN=iusf.b4770834&authtype=sso&custid=s3818721&site=eds-live&scope=site&custid=s3818721

Eckensberger, L. H. (2015). Integrating the emic (Indigenous) with the etic (Universal) – A case of squaring the circle or for adopting a culture inclusive action theory perspective. *Journal for the Theory of Social Behaviour*, *45*(1), 108–140. https://doi.org/10.1111/jtsb.12057

Edwards, J. (2006). Language revitalization and its discontents: An essay and review of saving languages: An introduction to language revitalization. In L. Grenoble & L. Whaley (Eds.). *Saving languages: An introduction to language revitalization* (pp. 101–120). Cambridge University Press.

Erickson, F., & Mohatt, C. (1982). Cultural organization and participation structures in two classrooms of Indian students. In G. Spindler (Ed.), *Doing the ethnography of schooling* (pp. 131–174). Holt, Rinehart & Winston.

Estrin, E. T., & Nelson-Barber, S. (1995). *Issues in cross-cultural assessment: American Indian and Alaska Native students*. Office of Educational Research and Improvement, U.S. Department of Education.

Fanon, F. (1952). *Black skin, white masks* (Repr.). Pluto Press. https://monoskop.org/images/a/a5/Fanon_Frantz_Black_Skin_White_Masks_1986.pdf

Farina, A. S. J., & Mancini, M. (2017). Evaluation of a multi-phase trauma-focused intervention with Latino youth. *Advances in Social Work*, *18*(1), 270–283. https://doi.org/10.18060/21296

Fisher, W. P., Jr., & Wilson, M. (2015). Building a Productive Trading Zone in Educational Assessment Research and Practice. *Pensamiento Educativo: Revista de Investigación Educacional Latinoamericana*, *52*(52), 55–78. https://doi.org/10.7764/PEL.52.2.2015.16

Foster, M., Lewis, J., & Onafowora, L. (2003). Anthropology, culture, and research on teaching and learning: Applying what we have learned to improve practice. *Teachers College Record*, *105*(2), 261–277. https://doi.org/10.1111/1467-9620.t01-1-00239

Freire, P. (2017). *Pedagogy of the oppressed*. Penguin Classics.

Garcia, A., Baethke, L., & Kaur, J. S. (2017). Lessons learned from Native C.I.R.C.L.E., a culturally specific resource. *Journal of Cancer Education*, *32*(4), 740–744. https://doi.org/10.1007/s13187-016-1001-x

United Nations Declaration on the Rights of Indigenous Peoples, A/RES/61/295 (02 October 2007), available from undocs.org/en/A/RES/61/295. (2007).

Gillborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: education, policy, 'Big Data' and principles for a critical race theory of statistics. *Race Ethnicity and Education*, *21*(2), 158–179. https://doi.org/10.1080/13613324.2017.1377417

Gion, C., McIntosh, K., & Smolkowski, K. (2018). Examination of American Indian/Alaska Native school discipline disproportionality using the vulnerable decision points approach. *Behavioral Disorders*, *44*(1), 40–52. https://doi.org/10.1177/0198742918773438

Gray, J. S., Brionez, J., Petros, T., & Gonzaga, K. T. (2019). Psychometric evaluation of depression measures with northern plains Indians. *American Journal of Orthopsychiatry*, *89*(4), 534–541. https://doi.org/10.1037/ort0000309

Grayshield, L., Rutherford, J. J., Salazar, S. B., Mihecoby, A. L., & Luna, L. L. (2015). Understanding and healing historical trauma: The perspectives of Native American elders. *Journal of Mental Health Counseling*, *37*(4), 295–307. https://doi.org/10.17744/mehc.37.4.02

Hegel, G. W. F. (2010). *The science of logic* (G. D. Giovanni, Trans.). Cambridge University Press.

Hitchcock, J. H., Nastasi, B. K., Dai, D. Y., Newman, J., Jayasena, A., Bernstein-Moore, R., Sarkar, S., & Varjas, K. (2005). Illustrating a mixed-method approach for validating culturally specific constructs. *Journal of School Psychology*, *43*, 259–278. https://doi.org/10.1016/j.jsp.2005.04.007

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. *The Journal of Negro Education*, *67*(3), 187–196.

Hood, S. (2018, May 13). *Project Nomenclature* [Personal communication].

Horkheimer, M. (2018). The state of contemporary social philosophy and the tasks of an institute for social research (1931). *Journal for Cultural Research*, *22*(2), 113–121. https://doi.org/10.1080/14797585.2018.1461354

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology*, *16*(2), 131–152.

Indian Education Act of 1979. (1982). *P.L. 95–561*, *25 U.S.C. §32.4 (1982)*.

Johnson, P. (2016). Indigenous knowledge within academia: Exploring the tensions that exist between Indigenous, decolonizing, and Nêhiyawak methodologies. *Totem: The University of Western Ontario Journal of Anthropology*, *24*(1), 4–61.

Johnson, S. ímlaʔxw M. K. (2017). Breathing life into new speakers: Nsyilxcn and Tlingit sequenced curriculum, direct acquisition, and assessments. *Canadian Modern Language Review*, *73*(2), 109–132.

Jordan, C. (1985). Translating culture: From ethnographic information to educational program. *Anthropology & Education Quarterly*, *16*(2), 105–123.

Kaltman, S., Hurtado de Mendoza, A., Serrano, A., & Gonzales, F. A. (2016). A mental health intervention strategy for low-income, trauma-exposed Latina immigrants in primary care: A preliminary study. *American Journal of Orthopsychiatry*, *86*(3), 345–354. https://doi.org/10.1037/ort0000157

Kataoka, S. H., Stein, B. D., Jaycox, L. H., Wong, M., Escudero, P., Tu, W., Zaragoza, C., & Fink, A. (2003). A school-based mental health program for traumatized Latino immigrant children. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*(3), 311–318. https://doi.org/10.1097/00004583-200303000-00011

Kinzel, C. A. (2015). *Developing culturally specific curriculum: Supporting Aboriginal early learners*. University of British Columbia.

Kroskrity, P. V. (2009). Language renewal as sites of language ideological struggle. The need for 'ideological clarification.' In J. A. Reyhner & L. Lockard (Eds.). *Indigenous language revitalization: Encouragement, guidance & lessons learned* (pp. 71–83). Northern Arizona University.

Ladson-Billings, G. (1995a). But that's just good teaching! The case for culturally relevant pedagogy. *Theory Into Practice*, *34*(3), 159–165. https://doi.org/10.1080/00405849509543675

Ladson-Billings, G. (1995b). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*(3), 465–491.

Ledlow, S. (1992). Is cultural discontinuity an adequate explanation for dropping out? *Journal of American Indian Education*, *31*(3), 21–36.

Lehrer, R. (2013). A learning progression emerges in a trading zone of professional community and identity. *WISDOMe Monographs*, *3*, 173–186.

Leigh-Osroosh, K. T., & Hutchison, B. (2019). Cultural identity silencing of Native Americans in education. *Race and Pedagogy Journal*, *4*(1), 1–33.

Lomawaima, K. T. (2000). Tribal sovereigns: Reframing research in American Indian education. *Harvard Educational Review*, *70*(1), 1–23. https://doi.org/10.17763/haer.70.1.b133t0976714n73r

Lomawaima, K. T., & McCarty, T. L. (2002). When Tribal sovereignty challenges democracy: American Indian education and the democratic ideal. *American Educational Research Journal*, *39*, 279–305. https://doi.org/10.3102/00028312039002279

Lovelace, S., & Wheeler, T. R. (2006). Cultural discontinuity between home and school language socialization patterns: Implications for teachers. *Education*, *172*(2), 303–309.

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing*, *18*(4), 351–372.

Maker, C. J. (2020). Culturally responsive assessments of spatial analytical skills and abilities: Development, field testing, and implementation. *Journal of Advanced Academics*, *31*(3), 234–253. https://doi.org/10.1177/1932202X20910697

Mallinson, T. (2024). Extending the justice-oriented, anti-racist framework for validity testing to the application of measurement theory in re(developing) rehabilitation assessments. In W. P. Fisher Jr. & L. R. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*. De Gruyter.

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement* (2nd ed.), Springer. https://link.springer.com/book/10.1007/978-3-031-22448-5

Maureen E., S. (2016). Special education pre-referrals in one public school serving Native American students. *Journal of American Indian Education*, *55*(2), 4–27. https://doi.org/10.5749/jamerindieduc.55.2.0004

McCarty, T. L. (2003). Revitalising Indigenous languages in homogenising times. *Comparative Education*, *39*(2), 147–163.

McCarty, T. L., & Lee, T. S. (2014). Critical culturally sustaining/revitalizing pedagogy and indigenous education sovereignty. *Harvard Educational Review*, *1*, 101. https://doi.org/10.17763/haer.84.1.q83746nl5pj34216

McCubbin, L., & Bennett, S. (2008). Cultural encapsulation. In F. T. Leong (Ed.). *Encyclopedia of counseling* (Vol. 3, pp. 1091–1091). Sage Publications, Inc. doi:10.4135/9781412963978.n352

McGroarty, M., Beck, A., & Butler, F. A. (1995). Policy issues in assessing Indigenous languages: A Navajo case. *Applied Linguistics*, *16*, 323–343.

Meek, B. A. (2007). Respecting the language of elders: Ideological shift and linguistic discontinuity in a Northern Athapascan community. *Journal of Linguistic Anthropology*, *17*(1), 23–43.

Meek, B. A. (2010). *We are our language: An ethnography of language revitalization in a Northern Athabascan community*. University of Arizona Press.

Morgan, R., & Freeman, L. (2009). The Healing of our people: Substance abuse and historical trauma. *Substance Use & Misuse*, *44*(1), 84–98. https://doi.org/10.1080/10826080802525678

Morris, R., Pae, H. K., Arrington, C., & Sevcik, R. (2006). The assessment challenge of Native American educational researchers. *Journal of American Indian Education*, *45*(3), 77–91.

Muir, N., & Bohr, Y. (2014). Contemporary Practice of Traditional Aboriginal Child Rearing: A Review. *First Peoples Child & Family Review*, *9*(1), 66–79.

Nastasi, B. K. (2000). School psychologists as health-care providers in the 21st century: Conceptual framework, professional identity, and professional practice. *School Psychology Review*, *29*(4), 540–554.

O'Brien, A. P., Boddy, J. M., & Hardy, D. J. (2007). Culturally specific process measures to improve mental health clinical practice: Indigenous focus. *Australian & New Zealand Journal of Psychiatry*, *41*(8), 667–674.

Pendrill, L. R. (2019). *Quality assured measurement: Unification across social and physical sciences*. Springer. https://link.springer.com/book/10.1007/978-3-030-28695-8

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders*. (3rd ed.), Allyn and Bacon.

Portes, P. R. (1999). Examining a cultural history puzzle. *American Educational Research Journal*, *36*(3), 489–507.

Ragoonaden, K., & Mueller, L. (2017). Culturally responsive pedagogy: Indigenizing curriculum. *Canadian Journal of Higher Education*, *47*(2), 22–46.

Santiago, C. D., Kataoka, S. H., Hu-Cordova, M., Alvarado-Goldberg, K., Maher, L. M., & Escudero, P. (2015). Preliminary evaluation of a family treatment component to augment a school-based intervention serving low-income families. *Journal of Emotional and Behavioral Disorders*, *23*(1), 28–39. https://doi.org/10.1177/1063426613503497

Seider, S., Graves, D., El-Amin, A., Soutter, M., Tamerat, J., Jennett, P., Clark, S., Malhotra, S., & Johannsen, J. (2018). Developing sociopolitical consciousness of race and social class inequality in adolescents attending progressive and no excuses urban secondary schools. *Applied Developmental Science*, *22*, 169–187.

Shea, H., Mosley-Howard, G. S., Baldwin, D., Ironstrack, G., Rousmaniere, K., & Schroer, J. E. (2019). Cultural revitalization as a restorative process to combat racial and cultural trauma and promote living well. *Cultural Diversity and Ethnic Minority Psychology*, *25*(4), 553–565. https://doi.org/10.1037/cdp0000250

Stake, R. (1973, October). *Program evaluation particularly responsive evaluation*. New Trends in Evaluation, Goteborg, Sweden.

Stake, R. (2011). Program evaluation particularly responsive evaluation. *Journal of MultiDisciplinary Evaluation*, *7*(15), 180–201.

Sul, D. A. (2021). *Indigenous assessment developers on elements of the disjuncture-response dialectic: A critical comparative case study* [Doctoral dissertation, University of San Francisco]. https://repository.usfca.edu/diss/571

Sul, D. A. (2019, March). *Reclaiming educational autonomy and minimizing measurement disjuncture through a culturally specific assessment development process*. [Paper presentation]. Culturally Responsive Evaluation and Assessment (CREA) Conference 2019, Chicago, IL. https://www.researchgate.net/publication/332275884_Reclaiming_educational_autonomy_and_minimizing_measurement_disjuncture_through_a_culturally_specific_assessment_development_process

Sul, D. A., & Domínguez, D. (2021). *The Development of the Latinx Immigration Trauma Construct: A Response to Measurement Disjuncture Using a Culturally Specific Assessment Model*. 1–20.

Telander, K. J. (2012). *An exploratory evaluation of a culturally specific model of psychological well-being for an African American population* [Doctoral dissertation, Loyola University Chicago]. http://ecommons.luc.edu/luc_diss/397

The Getting it Right Collaborative Group, Hackett, M. L., Teixeira-Pinto, A., Farnbach, S., Glozier, N., Skinner, T., Askew, D. A., Gee, G., Cass, A., & Brown, A. (2019). Getting it Right: Validating a culturally specific screening tool for depression (aPHQ-9) in Aboriginal and Torres Strait Islander Australians. *Medical Journal of Australia*, *211*(1), 24–30. https://doi.org/10.5694/mja2.50212

Thompkins, F., Goldblum, P., Lai, T., Hansell, T., Barclay, A., & Brown, L. M. (2020). A culturally specific mental health and spirituality approach for African Americans facing the COVID-19 pandemic. *Psychological Trauma: Theory, Research, Practice, and Policy*, *12*(5), 455–456. https://doi.org/10.1037/tra0000841

Thorndike, R. M., & Thorndike-Christ, T. (2009). *Measurement and evaluation in psychology and education* (8th ed.), Pearson.

Tillman, L. C. (2002). Culturally sensitive research approaches: An African-American perspective. *Educational Researcher*, *31*(9), 3–12. https://doi.org/10.3102/0013189X031009003

Tyler, K. M., Uqdah, A. L., Dillihunt, M. L., Beatty-Hazelbaker, R., Conner, T., Gadson, N., Henchy, A., Hughes, T., Mulder, S., Owens, E., Roan-Belle, C., Smith, L., & Stevens, R. (2008). Cultural discontinuity: Toward a quantitative investigation of a major hypothesis in education. *Educational Researcher*, *37*, 280–297.

van Meijl, T. (2006). Multiple identifications and the dialogical self: Urban Maori youngsters and the cultural renaissance. *Journal of the Royal Anthropological Institute*, *12*(4), 917–933. https://doi.org/10.1111/j.1467-9655.2006.00370.x

Vance, M. E., & Keele, B. (2013). Development and validation of the cervical cancer knowledge and beliefs of Appalachian Women Questionnaire. *Journal of Nursing Measurement*, *21*(3), 477–501. https://doi.org/10.1891/1061-3749.21.3.477

Venn, J. (1971). *Symbolic logic*. (2nd ed.), Chelsea Publishing Company.

Vining, C., Long, E., Inglebret, E., & Brendal, M. (2017). Speech-language assessment considerations for American Indian and Alaska Native children who are dual language learners. *Perspectives of the ASHA Special Interest Groups*, *2*(14), 29–40. https://doi.org/10.1044/persp2.SIG14.29

Walls, M. L., Whitbeck, L., & Armenta, B. (2016). A cautionary tale: Examining the interplay of culturally specific risk and resilience factors in Indigenous communities. *Clinical Psychological Science*, *4*(4), 732–743. https://doi.org/10.1177/2167702616645795

Warner, S. L. N. (1999). Kuleana: The right, responsibility, and authority of Indigenous peoples to speak and make decisions for themselves in language and cultural revitalization. *Anthropology & Education Quarterly*, *30*(1), 68–93. https://doi.org/10.1525/aeq.1999.30.1.68

Warschauer, M. (1998). Technology and indigenous language revitalization: Analyzing the experience of Hawai'i. *The Role of New Technologies in the Teaching of Second/Foreign Languages*, *1*, 139.

Whitfield, L. (2017). *Culturally specific interventions to support adolescent immigrant and refugee mental health* [Master's, St. Catherine University]. https://sophia.stkate.edu/msw_papers/811

Whitford, D. K. (2017). School discipline disproportionality: American Indian students in Special Education. *The Urban Review*, *49*(5), 693–706. https://doi.org/10.1007/s11256-017-0417-x

Wiesner, J. L. (2006). School climate interventions for Native American students: Minimizing cultural discontinuity in public schools. University of Wisconsin-Stout.

Wilson, M. R. (2023). *Constructing measures: An item response modeling approach* (2nd ed.), Routledge. https://doi.org/10.4324/9781410611697

Wixson, K. K., Dutro, E., & Athan, R. G. (2003). Chapter 3: The challenge of developing content standards. *Review of Research in Education*, *27*(1), 69–107. https://doi.org/10.3102/0091732X027001069

Wolfe, P. (1999). *Settler colonialism and the transformation of anthropology: The politics and poetics of an ethnographic event*. Cassell.

Wolfe, P. (2006). Settler colonialism and the elimination of the native. *Journal of Genocide Research*, *8*(4), 387–409. https://doi.org/10.1080/14623520601056240

Wyman, L., Marlow, P., Andrew, F. C., Miller, G. S., Nicholai, R. C., & Rearden, N. Y. (2010). Focusing on long-term language goals in challenging times: A Yup'ik example. *Journal of American Indian Education*, *49*, 28–49.

Yellow Bird, M. (1999). Indian, American Indian, and native Americas: Counterfeit identities. *Winds of Change: A Magazine for American Indian Education and Opportunity*, *14*(1). https://www.aistm.org/yellowbirdessay.htm

Contributors

Kenzo Asahi is an assistant professor at the School of Government at Pontificia Universidad Católica de Chile. His research interests lie in the intersection of program evaluation, urban economics, and labor economics.

Matt Barney is an award-winning organizational psychologist with over 25 years of diverse experience with multinationals like AT&T, Motorola, and Infosys. A serial entrepreneur, Dr. Barney founded XLNC, where he developed innovative Rasch guardrails for ethical AI measurement and unbiased evaluation. Prior to this, as the Founder of LeaderAmp, he was recognized with scientific awards for his AI from the Association of Test Publishers and the Society for Industrial-Organizational Psychology. With a consistent contribution to interdisciplinary science, he has secured four patents, published ten books, and authored over 250 publications and keynotes. He has served on the business affairs committee of not-for-profit scientific publisher Annual Reviews since 2014. Dr. Barney holds a PhD in Industrial-Organizational Psychology from the University of Tulsa and a BS in Psychology from the University of Wisconsin–Madison.

Feynman Barney is a third-year student at the University of California-Berkeley, where he studies Mechanical Engineering. Currently, he is contributing to Tesla's Cybertruck launch as a member of their Drive Unit team. Prior to his role at Tesla, Barney did research on superconductors at the Lawrence Berkeley National Lab's Applied Physics Division where he led automation efforts to measure changes in the superconductivity properties of physical materials.

Chris Bradley, PhD, is a research associate at the Wilmer Eye Institute of the Johns Hopkins University School of Medicine. He has a BA in mathematics from Columbia University and a PhD in sensory neuroscience from the University of Texas, Austin. His research spans basic vision science and psychometrics, as well as their applications to clinical research. His basic research includes advancing the retina-V1 model to predict the detectability of localized stimuli across the visual field from known properties of retinal ganglion cells. In psychometrics, he developed the method of successive dichotomizations and a latent variable extension of signal detection theory. In clinical research, he developed an improved method for estimating the accuracy of diagnosing glaucoma progression given optical coherence tomography measurements of retinal nerve fiber layer thickness and visual field measurements. He also developed the automated visual field test algorithm incorporated in the wearable Radius XR head-mounted display.

Yin Burgess obtained her PhD in Educational Research and Measurement from the University of South Carolina. During her academic career, she acquired valuable experience and developed a passion for assessment and psychometrics. She is currently employed as a psychometrician at the National Registry of EMTs where her focus is psychometric analyses and the test publication process for certification examinations. Prior to joining the National Registry, she worked at the Research, Evaluation, and Measurement Center at the University of South Carolina on various K-12 assessment programs and program evaluation projects. Her research interests include Rasch modeling, structural equation modeling, performance assessment, and survey design.

William P. Fisher, Jr. is recognized for contributions to measurement theory and practice spanning the full range from the philosophical to the applied in fields as diverse as education, mindfulness, survey research, organizational performance assessment, the clinical laboratory, and metrology. His entry in the 2011 World Standards Day paper competition won third prize, which is notable given the focus on engineering and natural science topics usually encountered in that context.

https://doi.org/10.1515/9783111036496-016

Jorge González is an associate professor at the Faculty of Mathematics, Pontificia Universidad Católica de Chile. His research interests include statistical modeling in social sciences, psychometrics, and educational measurement, with particular emphasis in test equating.

Hrishi Joshi is a consultant urological surgeon and honorary senior lecturer at the University Hospital of Wales since 2007. He completed his Doctor of Medicine in Bristol and Specialist urological training in East Anglia (Cambridge Deanery) 1999–2006. He has research interests, with international recognition, in the field of ureteral stents, urolithiasis, and outcome assessments. He serves on The Royal College Council in the Faculty and Associate Surgical Specialist lead roles. He mentors senior trainees as well as consultant colleagues across the UK and serves as a leader in the management of benign prostatic diseases, stones, and endourology services at the University Hospital of Wales.

Simon Karlsson holds an MSc in psychology and a BSc in political science. He currently serves as an analyst at RISE Research Institutes of Sweden within the Department of Measurement Science and Technology. His primary responsibilities involve conducting quality assessments of measurement tools designed to assess subjective and latent constructs through the application of psychometric analysis. Simon has experience in analyzing and improving measurement tools related to well-being, perceived safety, and the work environment, among other things. Additionally, he possesses expertise in questionnaire development, contributing to the enhancement of self-reported surveys.

Charalambos (Harry) Kollias is Research Director – Psychometrician in the Centre for Statistics (CfS) at the National Foundation for Educational Research (NFER) and directs/works on several international projects with responsibilities ranging from conducting, overseeing, and/or reviewing international large-scale survey data analysis and scale construction to evaluating international policy linking workshops. Harry has extensive experience in analyzing large-scale survey data and conducting (virtual) standard setting workshops to align assessment instruments to international frameworks such as the European Qualifications Framework (EQF), the Common European Framework of Reference for Languages (CEFR), and the Global Proficiency Framework (GPF). He has presented research at local and international conferences. He is highly experienced in a range of psychometric analysis, including test equating, item analysis and test-taker performance, and item bank calibration through Rasch Measurement Theory (RMT)/Item Response Theory (IRT) and Classical Test Theory (CTT). He has recently authored *Virtual Standard Setting: Setting Cut Scores* (Peter Lang).

D. Nantha Kumar is a medical undergraduate student in Cardiff University. She has a keen interest in surgical academia especially urology. Alongside co-authoring a peer-reviewed article she has presented at an international conference by the Association of Women Surgeons conference which led to her winning an award.

Dr. Trudy Mallinson is Associate Professor and Associate Dean for Research in Health Sciences in the School of Medicine and Health Sciences at The George Washington University. Her primary research interest is how better outcomes measurement can improve health care and inform health-care policy. She is particularly interested in how visualizing measurement data can help clinicians and patients can make better treatment decisions together. She advocates that clinical assessments should look and operate like rulers, so they can used that way: to measure a single dimension at a time, in order to compare real patient differences, regardless of who is using the assessment or who they are measuring. Her current research addresses a variety of rehabilitation measurement issues including measuring the recovery of consciousness in patients with severe traumatic brain injury, the standardization and calibration of functional performance assessments across rehabilitation and community settings, and the process of relationship-centered shared decision-making.

Robert W. Massof, PhD is Director Emeritus of the Lions Vision Research and Rehabilitation Center and Professor of Ophthalmology and Neuroscience at the Johns Hopkins University School of Medicine. His research is in the areas of vision psychophysics, psychometrics, and physiological optics applied to clinical problems in ophthalmology. He has 240 publications, 9 patents, and edited 2 books. He is a fellow of the Association for Research in Vision and Ophthalmology, Optica (Optical Society of America), and the American Academy of Optometry. His awards include Helen Keller Laureate (Helen Keller Foundation), Alfred W. Bressler Prize in Vision Science (Jewish Guild for the Blind), Pisart Vision Award (Lighthouse International), William Feinbloom Award (American Academy of Optometry), and Champion of Change Award (Obama White House).

Nadine LeBarron McBride is the Director of Psychometrics and Data Analytics for the National Registry of Emergency Medical Technicians. Prior to the National Registry, she served in psychometric and test development program management roles across a variety of certification and K-12 testing programs. She earned her BS in Psychology from the State University of New York at Albany and her PhD in Industrial/Organizational Psychology at Virginia Polytechnic Institute and State University. Her primary research interests lie in developing practical ways to gather and incorporate validity evidence throughout the test development cycle, supporting improvements in exam and item development and score interpretation.

Allison M. McCarthy, PhD, is Assistant Professor of Psychiatry and Behavioral Sciences with Vanderbilt School of Medicine and Core Faculty with the Center for Biomedical Ethics and Society at Vanderbilt University Medical Center. She completed her PhD in philosophy at The Ohio State University and a fellowship in clinical ethics at UCLA Health. Her research focuses on the conceptual and normative underpinnings of shared decision-making in clinical care and on the practical application of principles of shared decision-making to distinctive patient populations, specifically patients with intellectual and developmental disabilities and pregnant patients. Her work has been published in venues including NEJM, Kennedy Institute of Ethics Journal, and AJOB Neuroscience. She also serves as one of VUMC's full-time clinical ethicists, which supports shared decision-making between patients, families, and health-care professionals in ethically complex patient-care situations across all areas of adult and pediatric medicine.

Jeanette Melin holds a PhD in medical sciences. She is a researcher at RISE (Research Institutes of Sweden) within the Department of Measurement Science and Technology and PI for the RISE Platform Center for Category-Based Measures. She is also affiliated with the Swedish Defense University, working with cognitive measures for admission tests for basic military training. For the past decade, she has been conducting methodological research on self-reports and the alike. She has been engaged in questionnaire development and evaluation across a wide range of patient-reported outcomes and experience measures. She has also been involved in initiating a discussion for a future sustainable organization for measurement quality assurance of category-based measurements. Recently, she has been engaged as a researcher and work package leader in the EMPIR projects NeuroMET (15HLT04) and NeuroMET2 (18HLT09).

Linda Morell addresses critical issues in educational assessment and evaluation. She teaches in the Berkeley School of Education on evaluation theory, design, and methodology. She also conducts original research through the BEAR Center. As Co-PI on the NSF-funded project – Learning Progressions in Science: Analyzing & Deconstructing the Multiple Dimensions in Assessment – she investigated student understandings of scientific argumentation, the cross-cutting concept of patterns, and three content areas: natural resources, ecosystems, and the structure of matter. Morell also directs the IES-funded project – Developing & Testing Multi-Component Computer-Based Assessment Tasks for the Next Generation Science Standards – a project that brings together UC Berkeley, Stanford University, the SERP Institute, and the San Francisco Unified School District to connect practice and research.

Leslie Pendrill is a docent in experimental physics with most of his professional life devoted to metrology (i.e., quality-assured measurement). As Head of Research at the Swedish National Metrology Institute (1985–2012), he has played various leading roles, nationally and internationally, in metrological organizations, including chairmanship of EURAMET (www.euramet.org), the European Association of National Metrology Institutes (2009–2012). Pendrill's research and teaching interests range from frequency-stabilized lasers for primary length standards and gas density refractometry to methodologies for optimized measurement uncertainties. Since 2000 his foundational studies of human-based metrology investigate the applicability of traditional engineering of measurement systems centered on transducing instruments. Construct specification equations, particularly based on informational entropy when explaining task difficulty, have been formulated as anchors for interoperability of ordinal data and categorical classification.

Marcela Perticará is an associate professor at the School of Administration and Economics at Universidad Diego Portales, Chile. Her research interests include impact evaluation methodologies, gender disparities, childcare issues, and inequality.

Greg Sampson is a member of the psychometrics team at The National Registry of Emergency Medical Technicians. He holds academic credentials in quantitative research methods, data science, and analytics. His primary interests include statistical programming, operations research, and psychometric theory as these relate to high-stakes operational testing programs. He has worked across several assessment programs in pK-12, high-tech, vocational training, higher education, professional certification, and people analytics. Sampson earned his PhD from Oregon State University.

Ernesto San Martín is a full professor at the Faculty of Mathematics, Pontificia Universidad Católica de Chile and Invited Professor at the LIDAM/CORE, Université catholique de Louvain, Belgium. His research focuses on the modeling of social phenomena in politics, psychology, education, and the evaluation of public policies.

Brent A. Stevenor is an associate psychometrician for the National Registry of Emergency Medical Technicians. He earned his PhD in Industrial-Organizational Psychology from Bowling Green State University. His research interests include personality and individual differences, personnel selection and assessment, and psychometrics.

David Sul, EdD, is the Research Assistant Professor of Measurement at the University of the Virgin Islands and is the owner of Sul & Associates International, a professional measurement and evaluation firm. Sul, a critical psychometrician, works to educate the public on how the process of "measuring things that exist in the mind" can advance cross-cultural aspirations for autonomy, self-determination, and liberation. His current work focuses on the development of a generalized research strategy for the development of culturally specific assessments.

Hanna Svensson holds an MSc in Applied Physics and Electrical Engineering and serves as a dedicated researcher at RISE (Research Institutes of Sweden) within the department of Measurement Science and Technology. Her focus lies in ensuring the quality of categorical-based measurements and contributing to the realm of digitalization. Hanna has actively engaged in the development and analysis of fit-for-purpose measurement tools for municipalities, showcasing her commitment to practical applications. Over the last decade, she has been developing algorithms for advanced control systems, such as automated driving, wireless transmissions, and sensing of unobservable states.

Sean Tan is a doctoral candidate at the University of California, Berkeley. His research interests are in the measurement of cross-disciplinary and soft skills. Prior to that, he was a lead research specialist at the Singapore Ministry of Education, where he worked on international benchmarking and research studies, including the Teaching and Learning International Survey (TALIS), Programme for International Student Assessment (PISA), and Assessment and Teaching of 21st Century Skills (ATC21S). Sean obtained his bachelor's degree in Chemistry from the University of Cambridge and a master's in Educational Research, Measurement and Evaluation from Boston College. He has also taught A-level Chemistry and served in the school management team in Singapore schools.

Inés M. Varas is an adjunct professor at the Faculty of Mathematics, Pontificia Universidad Católica de Chile. Her research interests include statistical modeling on social sciences and biostatistics.

Judy R. Wilkerson is a Professor of Assessment, Evaluation, and Research in the College of Education at Florida Gulf Coast University, where she teaches undergraduate and graduate students. She earned her PhD at the University of South Florida in Measurement and Evaluation and focuses her teaching, research, and service on three related themes: (1) the credible (valid, reliable, and fair) assessment of the cognitive and affective domains of all learners; (2) the pragmatic applications of measurement and evaluation theory in the teacher education and the general higher education communities; and (3) the evaluative function of accreditation from a theoretical and practical standpoint. She has consulted extensively at international, national, and state levels in the areas of professional and regional accreditation.

Mark Wilson is Professor in the Berkeley School of Education at the University of California, Berkeley. He is a past President of the Psychometric Society and of the National Council on Measurement in Education, a Fellow of the American Educational Research Association, and is the recipient of multiple career recognition awards from professional associations. His interests focus on measurement and applied statistics. His work spans a range of issues in measurement and assessment from the development of new statistical models for analyzing measurement data, to the development of new assessments in subject matter areas such as science education, patient-reported outcomes, and child development, and to policy issues in the use of assessment data in accountability systems.

Jacob Wisén holds an MSc in International Social and Public Policy. He is a former project manager at RISE (Research Institutes of Sweden) in the Department of Measurement Science and Technology, where he has lead projects aimed at developing fit-for-purpose measurement tools for municipalities and civil society organizations. He is currently engaged in the development of community building models and evaluation tools at the Urban Development department at Stena Property.

Index

https://doi.org/10.1515/9783111036496-017

De Gruyter Series in Measurement Sciences (DGSMS)

Already published in the series

Metrological Infrastructure
Beat Jeckelmann, Robert Edelmaier (Eds.)
ISBN 978-3-11-071568-2, e-ISBN (PDF) 978-3-11-071583-5,
e-ISBN (EPUB) 978-3-11-071590-3

www.degruyter.com