Springer Series in Measurement Science and Technology

William P. Fisher, Jr.
Stefan J. Cano *Editors*

Person-Centered Outcome Metrology

Principles and Applications for High Stakes Decision Making

OPEN ACCESS

Springer

Springer Series in Measurement Science and Technology

The Springer Series in Measurement Science and Technology comprehensively covers the science and technology of measurement, addressing all aspects of the subject from the fundamental principles through to the state-of-the-art in applied and industrial metrology, as well as in the social sciences. Volumes published in the series cover theoretical developments, experimental techniques and measurement best practice, devices and technology, data analysis, uncertainty, and standards, with application to physics, chemistry, materials science, engineering and the life and social sciences.

William P. Fisher Jr. • Stefan J. Cano
Editors

Person-Centered Outcome Metrology

Principles and Applications for High Stakes Decision Making

Springer

Editors
William P. Fisher Jr.
University of California
Berkeley, CA, USA

Stefan J. Cano
Modus Outcomes (a Division of Thread)
Cheltenham, UK

This book is an open access publication. This title is freely available in an open access edition with generous support from the Library of the University of California, Berkeley.
ISSN 2198-7807 ISSN 2198-7815 (electronic)
Springer Series in Measurement Science and Technology
ISBN 978-3-031-07464-6 ISBN 978-3-031-07465-3 (eBook)
https://doi.org/10.1007/978-3-031-07465-3

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

This work was inspired by the insights and collegiality of Benjamin D. Wright and A. Jackson Stenner. Both made fundamental and lasting contributions to measurement research and development that will continue to guide theory and practice for decades to come.

Preface

The chapters in this book document the valuable work being done by researchers who have persisted in a radically different *metrological* vision of the role to be played by quantitative methods in healthcare outcomes management. These researchers have succeeded in bringing forth mutually complementary and coherent perspectives that are paradigmatically set apart from mainstream practice. Where the dominant statistical paradigm prioritizes the objectivity of data as the basis for credible generalizable conclusions, the authors of the chapters in this book instead prioritize the metrological objectivity of a quality-assured unit quantity as a trustworthy basis for communicating in common languages. Where the mainstream prioritizes centralized planning, data gathering and analysis, and statistical hypothesis testing as signature hallmarks of quantitative methodology, the authors of the chapters in this book instead complement data with distributed networks of instruments and predictive theories.

Instead of merely numeric results signifying often incomparable or unknown consequences across data sets, the chapters in this book focus on enabling individuals to tell personally meaningful stories of healing, development, and improved outcomes. Metrologically speaking, trust is less a function of isolated and disconnected facts, however objective they may be, than it is a function of the demonstrated explanatory power of theoretical predictions and the reproducibility of experimental tests. The chapters in this book provide insights into the principles involved in these tests and predictions, with multiple examples.

Though delayed by the undeniable impositions of the pandemic, and though only about two thirds of the planned chapters are included here in the final product, this book represents what we feel to be the beginning of new directions in healthcare outcomes measurement and management. Several of the authors unable to participate in this volume are eager to contribute to a second effort of this kind, as are other colleagues also working in this area. We hope in due course to produce not only another collection in this vein but to also participate in the emergence of new professional associations, journals, textbooks, and standards groups in the coming years.

That is, we recognize the necessary and complementary roles that must be played by literary, social, and material technologies if fundamental changes in healthcare are ever to be achieved. The mere objective existence of facts has never been sufficient to the tasks of organizing sciences or economies. Rather, human interests in some facts and not others must cohere in forms of social and political organization that make objectively reproducible phenomena communicable in shared, trusted systems. The chapters brought together here represent the barest beginning of efforts aimed at creating such systems.

Berkeley, CA, USA — William P. Fisher Jr.
Cheltenham, UK — Stefan J. Cano

Acknowledgments

William Fisher thanks his friend and co-editor, Stefan, for initiating this project and for the many years of engaging dialogue. William also thanks his colleagues at the Research Institutes of Sweden (RISE, Gothenburg), the Teaching Institute for Excellence in STEM (TIES, Cleveland, Ohio), and at the University of California (Berkeley) for ongoing dialogues and stimulating exchanges of ideas. Particular value has been contributed by Leslie Pendrill and Jeanette Melin at RISE, Jan Morrison at TIES, and Mark Wilson in Berkeley. We don't always agree on the specifics, but finding the differences that make a difference continues to guide the way.

Stefan Cano would like to thank his mentor, friend, and co-editor William for opening doors to a new world of possibilities in person-centered outcome metrology. Stefan also thanks Leslie and Jeanette at RISE for their collaboration and collegiality in the NeuroMet project. Finally, Stefan is grateful to the following people for making his 15-year journey into the world of metrology such a rich and engaging experience: David Andrich (University of Western Australia), Thomas Salzberger (University of Vienna), Andrea Pusic (Harvard Medical School), Anne Klassen (McMaster University), John Browne (University of Cork), Patrick Marquis (Modus Outcomes, USA), and Antoine Regnault (Modus Outcomes, France).

Contents

Chapter 1
Ideas and Methods in Person-Centered Outcome Metrology

William P. Fisher Jr. and Stefan J. Cano

Abstract Broadly stated, this book makes the case for a different way of thinking about how to measure and manage person-centered outcomes in health care. The basic contrast is between statistical and metrological definitions of measurement. The mainstream statistical tradition focuses attention on numbers in centrally planned and executed data analyses, while metrology focuses on distributing meaningfully interpretable instruments throughout networks of end users. The former approaches impose group-level statistics from the top down in homogenizing ways. The latter tracks emergent patterns from the bottom up, feeding them back to end users in custom tailored applications, whose decisions and behaviors are coordinated by means of shared languages. New forms of information and knowledge necessitate new forms of social organization to create them and put them to use. The chapters in this book describe the analytic, design, and organizational methods that have the potential to open up exciting new possibilities for systematic and broad scale improvements in health care outcomes.

Keywords Statistical modeling · Measurement modeling · Complexity · Distributed thinking

W. P. Fisher Jr. (✉)
BEAR Center, Graduate School of Education, University of California, Berkeley, CA, USA

Research Institutes of Sweden, Gothenburg, Sweden

Living Capital Metrics LLC, Sausalito, CA, USA
e-mail: wpfisherjr@livingcapitalmetrics.com

S. J. Cano
Modus Outcomes (a Division of Thread), Cheltenham, UK

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3_1

1.1 Introduction

The domain of “person centered outcomes” is an evolving array of ideas, tools, and practices. In this book, we use person centered outcomes (PCOs) as an umbrella term to encompass key stakeholders’ (i.e., the recipient,[1] caregiver or provider of healthcare) assessments, ratings, beliefs, opinions, experience or satisfaction concerning medical/surgical interventions (including clinical practice, research, trials). PCO instruments (e.g., rating scales, ability tests, biometric equipment, wearables) purport to quantify health, health-related quality of life and other latent health constructs, such as pain, mood, and function. These can also be used to quantify quality of healthcare. PCO instruments play an increasingly central role in evidence-based medicine [93, 107, 108].

Used alone, or in tandem with surrogate data (e.g., analyzed in the laboratory), PCO data offer the opportunity for more meaningful and interpretable *individualized* measures of patient outcomes. Custom-tailored PCO reports do not entail either superficially comparable numbers or completely disconnected details. PCO data, instruments, and theory have repeatedly—across multiple clinical situations—proven themselves dependable foundations for meaningful common languages and shared metrics that speak directly to care recipients, caregivers, and healthcare providers, researchers, and policy makers [30, 39, 59].

Meaningful PCO measures map the natural courses of disease, healing, degenerative conditions, learning, development, and growth in quality-assured common units with clearly stated uncertainties, to guide treatment decisions tailored to the unique situations of different patients. Most current approaches to person-centeredness are limited, in that they typically do not follow through from the stated intentions of focusing on people (patients, employees, clients, customers, suppliers, students, etc.) to fulfillment in practice [20, 33]. The crux of the matter concerns the difference between modernizing and ecologizing, which refers to prioritizing the objectivity of the data in disconnected statistical modeling, versus prioritizing networks of actors who agree on the objectivity of a unit quantity that retains its properties across samples and instruments [53, 76].

Therefore, it is important that we can clearly articulate how scientifically calibrated and metrologically distributed metrics—measurement *systems*—fulfill the meaning of person-centeredness in strikingly new and widely unanticipated ways. We offer seven suggestions.

- First, instead of burying the structure and meaning of patients’ expressions of their experiences in sum scores and ordinal statistics, we advocate using response data and explanatory models [35, 87, 104] to calibrate quality-assured instruments expressing that experience in substantively interpretable interval quantitative terms that are uniformly comparable everywhere [84, 85, 93, 94, 116, 118].

[1] For brevity, we use the term ‘patient’ as a shorthand for the recipient of health care or treatment, although we acknowledge this is quickly becoming an outdated term.

- Second, instead of perpetuating the failed assumption that objective reality somehow automatically propagates itself into shared languages and common metrics for free, we acknowledge and leverage networks of actors who agree on the objectivity of repeatable and reproducible structural invariances, and who collaborate in bringing those invariances into distributed measurement systems, usually at great expense, but also with proportionate returns on the investments [5, 13, 17, 37, 43, 45, 47, 48, 54, 68, 69, 72, 77, 120].
- Third, instead of using vaguely defined terms and policies to promote patient engagement and the improved outcomes that follow from informed patient involvement, we advocate defining it by mapping it, calibrating it, explaining it, and individualizing the navigation of it [16, 86, 105, 112, 132].
- Fourth, instead of assuming data are inherently valid and meaningful, we advocate theoretical explanations of patient experiences that support a qualitative narrative accounting for variation [35, 87, 104]; this sets up a new level of defensibility, not solely reliant on any given provider of healthcare's skills and experience.
- Fifth, instead of reifying unidimensionality in a rigid and uncompromising way, we take the pragmatic idealist perspective of using empirically and theoretically validated standards to illuminate differences that make a difference, and, conversely, tapping even small degrees of correlation between different dimensions for the information available [3, 113, 119].
- Sixth, instead of siphoning off data into research and management reports incapable of affecting the care of the individual patients involved, we advocate immediately feeding back at the point of care coherent [53, 111] contextualized and structured diagnostic reports; i.e., self-scoring forms and "kidmaps" which we may call "PatientMaps", "ClientMaps", or "PersonMaps" [12, 18, 26, 27, 50, 79, 80, 86, 111, 114, 115, 131, 132].
- Seventh, instead of assuming that statistical averages of ordinal scores are adequate to the needs of individual patient care, and instead of assuming even that logit measures and uncertainties are capable of summarizing everything important about an individual patient experience, we advocate displaying patterns of individual ratings illustrating diagnostically relevant special strengths and weaknesses; by acknowledging the multilevel semiotic complexity of all signification in language in this way, we recognize the nature of measured constructs as boundary objects "plastic enough to be adaptable across multiple viewpoints, yet maintain continuity of identity" [45, 47, 54, 101, p. 243].

An additionally useful reporting application would associate anomalous observations with diagnostically informative statistics drawn from historical data on similar patients with similar response patterns, conditions, co-morbidities, genetic propensities, etc. Guttman scalograms [63], for instance, used in conjunction with model fit statistics, reveal stochastic patterns in individual responses [78] predicting signature sequences of diagnostically informative departures from expectation [34, 61].

Metrological quality assurance is essential if reliable decisions about diagnosis, treatment and rehabilitation are to be made consistently throughout a healthcare system, with continuous improvement [25]. This goes far beyond the well-trodden path of debates about data analysis or model choice, which have played out ad nauseum, accomplishing little more than endless arguments over arbitrary methodological criteria. A description of the situation dating to over 20 years ago remains as true now as it was then. The fundamental oversight in person-centered health care outcome management is that, in addition to the problem of model choice,

> 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 (intra-laboratory 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. [43, p. 529]

Objective metrological comparability ('traceability') and declared measurement uncertainty leverage patterns that have been repeatedly reproduced for decades across patients and instruments, and that cohere into a common language [53, 111]. A possible way forward involves a synthesis of metrology, psychometrics, and philosophy that involves four cornerstones.

First, it is essential to root measured constructs and unit quantities succinctly and decisively in the ways they:

- are structured as scientific models in the form of Maxwell's equations, following Rasch [44, 48, 96, pp. 110–115];
- extend in new ways everyday language's roots in the metaphoric process [49], following Maxwell's method of analogy in his exposition of how "every metaphor is the tip of a submerged model" [14, 15, p. 30] and
- extend everyday thinking into new sciences in the manner described by Nersessian's [89] study of Maxwell's method of analogy [44, 45, 48].

Second, it is furthermore also essential to show and assert that measured constructs and unit quantities:

- are defined by the populations of persons and items manifesting the construct;
- are substantiated empirically in terms of samples of persons and items drawn from those populations that are rigorously representative of them; and
- are explained theoretically in terms of predictive models structuring experimental tests of cognitive, behavioral, and structural processes.

Third, from this it follows that:

- reference standard units and associated uncertainties will be set up as formal constants open to testing, refinement, and reproduction anywhere, anytime, by anyone;

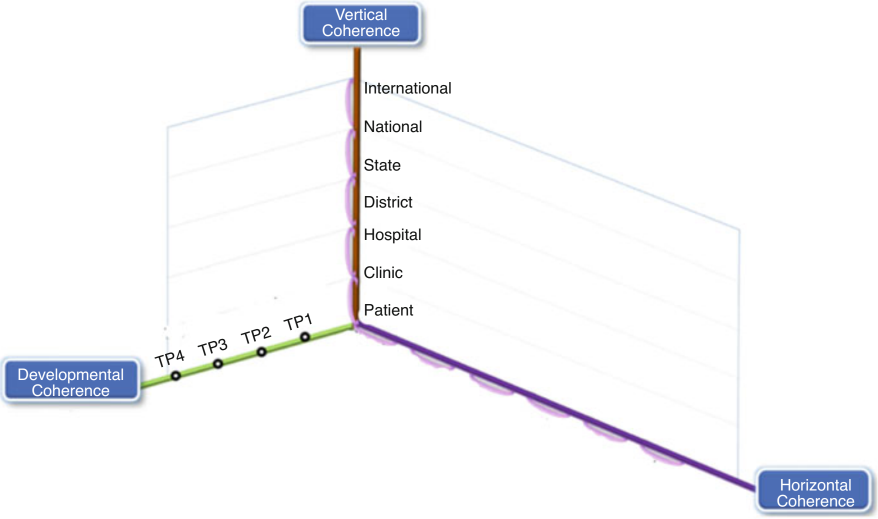


Fig. 1.1 Developmental, horizontal, and vertical coherent measurement dimensions. (Modified from Fisher et al. [51])

- criteria for sample definitions, instrument administration, data security, etc. will have to be developed and adopted via consensus processes; and
- local reference standard laboratories will be charged with reproducing the unit from standard samples and from theory, to within a relevant range of uncertainty, and maintaining it in clinical practice and research applications.

Fourth, expanding and clarifying these points:

- day-to-day measures will not be estimated via data analysis, but will instead be read from the calibrated instrument and will be reported in varying ways depending on the application:
  - individualized ‘kidmaps’ reporting specific responses;
  - measurements in the unit quantity and uncertainty; and
  - aggregate comparisons over time, horizontally across clinics and providers, and vertically within an organization, system, or region (see Fig. 1.1);
- quality assurance processes in the reference labs and the standard setting lab will document legally binding conformity with procedures;
- stakeholder participation in every area of activity and completely transparent openness to every kind of critical input will be essential; and
- we would warmly welcome every conceivable empirical and/or theoretical challenge because the contestability of comparable results is a hallmark precursor of scientific progress that has to date been counterproductively excluded and omitted from the methods of outcome modelling and measurement in health care and other fields.

The urgent need for a new focus is the key motivating factor for this edited volume. In this unique collection, we explore the synthesis of metrology, psychometrics, philosophy, and clinical management to support the global comparability and equivalence of measurement results in PCO measurement. The target audience for this book is any and all key stakeholders interested in person-centered care including policy makers, clinicians, pharmaceutical industry representatives, metrologists, and health researchers.

This book includes a unique collection of works from world-recognized experts, researchers and thought leaders in PCO research. The two sections of this volume explore the potential benefits of *moving towards* a PCO metrological framework across clinical practice and research, methodology and theory to provide solutions including:

- addressing the lack of units in patient centered outcome measurement through recourse to mathematical models devised to define meaningful, invariant, and additive units of measurement with known uncertainties;
- establishing coordinated international networks of key stakeholders guided by five principles (i.e., collaboration, alignment, integration, innovation and communication); and
- better use of technology leveraging measurement through item banks linking PCO reports via common items, common patients, or specification equations based in strong explanatory theory.

1.2 The Chapters

Section one includes five chapters covering person centered research and clinical practice. In her clinician's guide to performance outcome measurements, Anna Mayhew provides excellent insight as a clinical evaluator and researcher as to the role of the Rasch model in maximizing the use and interpretability of the North Star Ambulatory Assessment in better understanding the progression of Duchenne muscular dystrophy. Continuing this theme, Diane Allen and Sang Pak provide a clinical perspective as to what drives PCO measurement strategies in patient management.

We then turn to ophthalmology in two research programs. The first, from Maureen Powers and William P. Fisher, Jr. describes how physical and psychological measurements of vision combine into a model of functional binocular vision; this psychophysical combination of biological, survey, and reading test data demonstrates how data from different domains can be integrated in a common theoretical and applied context. The next chapter, from Bob Massof and Chris Bradley, describes the evolution of a long-standing program for low vision rehabilitation, which exploits item banking and computer adaptive testing. They propose a strategy for measuring patient preferences to incorporate in benefit-risk assessments of new ophthalmic devices and procedures. Finally, Sarah Smith describes the importance of quantitative and qualitative enquiry, against the backdrop of calibrated rating scales, providing the perspective of a health services researcher working in the field of dementia, at the national level.

In Section two, we move to fundamentals and applications. The section begins with John Michael Linacre's reflections on equating measurement scales via alternative estimation methods; conceptually similar scales are aligned so that the derived measures become independent of the specifics of the situation on which they are based, with the concomitant result that theoretical differences between supposedly superior and inferior estimation methods turn out to have little or no practical consequences. David Andrich and Dragana Surla tackle the same subject from the perspective of one estimation method, but with the goal of having a common unit referenced to a common origin and where the focus is on making decisions at the person level. Thomas Salzberger takes this one step further by providing an example from the measurement of substance dependence, making the argument for traceability in social measurement via the co-calibration of different instruments in a common metric.

Jeanette Melin and Leslie Pendrill provide two chapters, which take the conversation about co-calibration an additional step further, returning to the subject of dementia. First, the authors describe a research program which elaborates the role of construct specification equations and entropy to better understand the measurement of memory through ability tests. The subsequent chapter makes the link to quality assurance in PCO measurement by describing the potential for establishing metrological references in fields such as person-centered care in the form of "recipes" analogous to certified reference materials or procedures in analytical chemistry and materials science. Finally, William Fisher grounds the contents of this book in a philosophical framework extending everyday thinking in new directions that offer hope for achieving previously unattained levels of efficacy in health care improvement efforts.

1.3 Acknowledging and Incorporating Complexity

We expect the reader will recognize that there are potential inconsistencies and even disagreements across the chapters. We fully acknowledge these, and would respond that, though matters are very far from being resolved in any kind of a settled way, there are productive, constructive, and pragmatic reasons for considering a metrological point of view on the role of measurement in health care's person-centered quality improvement efforts.

Of particular importance among these reasons are the ways in which metrology undercuts the "culture wars" and the futile modernist-postmodernist debates, doing so by taking the focus off the relative priorities of theory vs observation [73, 74]. In Golinski's [60, p. 35] words, "Practices of translation, replication, and metrology have taken the place of the universality that used to be assumed as an attribute of singular science." Alternatively, in Haraway's [64, pp. 439–440] terms,

"…embedded relationality is the prophylaxis for both relativism and transcendence." That is, the universality of scientific laws cannot be demonstrated absent instrumentation and those laws cannot be communicated without a common language; nor can the observed data's disparate local dependencies make any sense in relation to anything if there is no metric or linguistic standard to provide a medium of comparison.

Both modern and postmodern perspectives must inevitably make use of shared standards, suggesting a third alternative focused on the shared media of communications standards and metrologically traceable instruments. Latour's [72, pp. 247–257] extended consideration of the roles of metrology is foundational. Latour [73, 74] characterizes this third alternative as amodern, and Dewey [36, p. 277] similarly arrives at a compatible unmodern perspective, saying that "…every science and every highly developed technology is a systematic network of connected facts and operations." Galison [57] considers the modern focus on transcendental universals as positivist, the postmodern emphasis on relativism as antipositivist, and the unmodern inclusion of the instrument as postpositivist. A large and growing literature in science and technology studies pursues the implications of instruments and standards for understanding the progress of science [1, 5, 13, 19, 21, 37, 67, 90, 109].

Galison [58, p. 143] offers an "open-ended model" of how different communities of research and practice interrelate. This perspective allows:

- partial autonomy to each community at their level of complexity:
  - experimentation's focus on concrete observable data,
  - instrumentation's focus on abstract communications standards, and
  - theory's focus on formal models, laws, and predictive theories; and
- "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" [57, p. 143].

A significant consequence of this open-ended model in physics is, Galison [57, pp. 46–47] points out, that

> …between the scientific subcultures of theory and experiment, or even between different traditions of instrument making or different subcultures of theorizing, there can be exchanges (co-ordinations), worked out in exquisite detail, without global agreement. Theorists and experimenters, for example, can hammer out an agreement that a particular track configuration found on a nuclear emulsion should be identified with an electron and yet hold irreconcilable views about the properties of the electron, or about philosophical interpretations of quantum field theory, or about the properties of films.
>
> The work that goes into creating, contesting, and sustaining local coordination is, I would argue, at the core of how local knowledge becomes widely accepted. At first blush, representing meaning as locally convergent and globally divergent seems paradoxical. On one hand, one might think that meaning could be given sentence by sentence. In this case the global sense of a language would be the arithmetical sum of the meaning given in each of its particular sentences. On the other hand, the holist would say that the meaning of any particular utterance is only given through the language in its totality. There is a third alternative, namely, that people have and exploit an ability to restrict and alter meanings

> in such a way as to create local senses of terms that speakers of both parent languages recognize as intermediate between the two. The resulting pidgin or creole is neither absolutely dependent on nor absolutely independent of global meanings.

What Galison describes here is the state of being suspended in language, semiotically, where abstract semantic standards mediate the negotiation of unrealistic conceptual ideals and unique, concrete local circumstances. This theme also emerges in the work of S. L. Star under the heading of the boundary object [19, 101–103], and in Woolley and Fuchs' [121] contention that healthy scientific fields must incorporate both divergent and convergent thinking.

An obvious point of productive disagreement in this vein emerges in the chapter by Massof and Bradley, with their "heretical" statements about expecting and accepting failures of invariance in their low vision rehabilitation outcomes measurement and management system. Differential item functioning and differential person functioning take on a new significance when theory explains the structural invariance incorporated in a measurement standard, and item location estimates have been stable across thousands or even millions of cases. A variation on this point is raised by Allen and Pak in the section in their chapter on the tensions between standardization and personalization. Here, local failures of invariance become actionable and relevant bits of information clinicians and others need to know about if they are to be able to formulate effective interventions.

It is part of the nature of a boundary object to accept those concrete levels of manifestations of unique patterns in the diagnosis-specific ways described by Allen and Pak, and by Massof and Bradley. As Star [101, p. 251] put it,

> …boundary objects…are a major method of solving heterogenous problems. Boundary objects are objects that are both plastic enough to adapt to local needs and constraints of the several parties employing them, yet robust enough to maintain a common identity across sites. They are weakly structured in common use, and become strongly structured in individual-site use.

Star and Ruhleder [103, p. 128] similarly say, "only those applications which simultaneously take into account both the formal, computational level and the informal, workplace/cultural level are successful." As is suggested in the chapter by Fisher, might the ongoing failures of person-centered quality improvement efforts listed by Berwick and Cassel [11] derive from inattention to the nature of boundary objects?

In that same vein, a more pointed instance of the heterogeneity of perspectives implied by boundary objects emerges in the longstanding debates between item response theory (IRT) and Rasch model advocates [4, 31, 42]. The arguments here focus on the descriptive value of statistical models obtaining the lowest p-values in significance tests, vs the prescriptive value of scientific models providing narrative explanations of variation and information, with additional indications as to how instruments and sampling procedures might be improved. These purposes are not, of course, always pursued in mutually opposed ways, and in current practice, both of them typically assume measurement to be achieved primarily via centrally planned and executed data analyses, not via the distributed metrology of calibrated instruments advocated here.

But in this metrological context, the “Rasch debate” [42] is defused. Data analysis certainly has an essential place in science, even if it should not be the primary determining focus of measurement. Boundary objects align with a kind of pragmatic idealism that recognizes there are communities, times, and places in which each of the different levels of complexity represented by data, instruments, and theory is valid and legitimate. There are just as many needs for locally intensive investigations of idiosyncratic data variations as there are for interconnected instrument standards and globally distributed explanatory theories.

But there are different ways of approaching local failures of invariance. It is essential to not confuse levels of complexity [45, 47]. Data analyses of all kinds can be productively pursued in hierarchically nested contexts bound by consensus standards structuring broad communications. But local exceptions to the rule that do not clearly invalidate the empirical and theoretical bases of item calibrations should no longer be allowed to compromise the comparability of measurements. There is nothing radical or new in saying this. It has long been recognized that “The progress of science largely depends on this power of realizing events before they occur,” that “laws are the instruments of science, not its aim,” and that “the whole value. . .of any law is that it enables us to discover exceptions” [32, pp. 400, 428, 430]. Instead of conceiving measurement primarily in terms of statistically modeled multivariate interactions, a larger role needs to be made for scientific modeling of univariate dimensions, as these are the means by which metrology creates labor-saving “economies of thought” [7, 32, pp. 428–429, 55, 56, 83, pp. 481–495].

Butterfield [24, pp. 16–17, 25–26, 96–98] notes that, in the history of science, observations do not accumulate into patterns recognized as lawful; instead, science advances as new ways of projecting useful geometric idealizations are worked out. Measurement structures linear geometries affording conceptual grasps of concrete phenomena by positing the hypothesis that something varies in a way that might be meaningfully quantified. Kuhn [71, p. 219; original emphasis] makes the point, saying,

> *The road from scientific law to scientific measurement can rarely be traveled in the reverse direction.* To discover quantitative regularity, one must normally know what regularity one is seeking and one's instruments must be designed accordingly; even then nature may not yield consistent or generalizable results without a struggle.

Practical metrological implementations of the results of measurement modeling exercises that begin by specifying the structure of lawful regularities, as in the use of Rasch’s models for measurement, require agreements on standardized criteria for knowing when and if comparability is substantively threatened. Efforts in this direction are being taken up, for instance, in the European NeuroMet project [39].

Taken out of context, the unfortunate effect of compromising the invariance of the unit quantity in the application of IRT models with multiple item parameters is that data are described to death. The value of identified models [98, 100], such as Rasch’s, concerns the practical implications of structural invariances for policy and programs. Rasch and Thurstone are acknowledged for their contributions in “fruitful” discussions concerning the development of the concept of identified models, those that require structural invariances reproducible across samples and instruments [70, p. 165]. Referred to by one of Rasch's mentors, Frisch, as "autonomy" [2], this quality in cross-data patterns is characteristic of a class of models necessary to learning generalizable lessons informing practical capacities for predicting the future [48]. Over-parameterized models, in contrast, may achieve statistical significance only at the expense of practical significance, such that the particular relationships obtained in a given data set are so closely specified that they are useless for anticipating future data [6, 23, 38, p. 211, 81, p. 22, 99, 110, p. 235; 123, 127].

By sweeping unexpected responses into data-dependent item parameters, exceptions to the rule are concealed in summary statistics and hidden from end users who might otherwise be able to make use of them in the manner described in the chapters by Allen and Pak, and by Massof and Bradley. But the *disclosure* of anomalies is well-established as a primary function of measurement. Rasch [96, pp. 10, 124], Kuhn [71, p. 205], and Cook [32, p. 431] all illustrate this point using the example of the discovery of Neptune from perturbations in the orbit of Uranus. Burdick and Stenner [22] concur, noting that IRT models put analysts in a position akin to Ptolemaic astronomers, whose descriptive approach to planetary orbits was more accurate than could be achieved using Newton's laws. What if astronomy had stuck with the Ptolemaic methods instead of adopting new ones based on physical theory? Ptolemaic astronomers can be imagined saying, "Forget about those perturbations in the orbit of Uranus. Our model accounts for them wonderfully." If astronomy as a field had accepted that position, instead of insisting on the prescriptive Newtonian model, then Neptune never could have been discovered by estimating the position and mass of an object responsible for perturbations of the magnitude observed in Uranus' orbit.

Though this process of being able to perceive exceptions only in relation to a standard may seem to be a highly technical feature of mathematical science, it is but an instance of the fact recognized by Plato in *The Republic* (523b–c) that "experiences that do not provoke thought are those that do not at the same time issue a contradictory perception." That is, we do not really think much at all about experiences meeting our expectations. The labor-saving economy of thought, created when language pre-thinks the world for us, removes the need to bother with irrelevant details. Scientific instruments extend this economy by embodying invariant meaning structures that predict the form of new data.

But in contrast to these more typical situations, innovative ideas and thoughtful considerations tend to follow from observations that make one think, "That's odd...." And so it happened with the discovery of penicillin when a lab culture died, the discovery of x-rays when a lead plate was misplaced in a lab, of vulcanization when liquid rubber was accidentally left on a hot stove, of post-it notes when an experimental glue did not stick, etc.

Nature is revealed by means of exceptions that are often products of serendipitous accidents [97]. Anomalous observations are answers to questions that have not been asked. Because attention is focused on conceptually salient matters already incorporated in the linguistic economy of thought, most unexpected observations are ignored as mere noise, as nuisance parameters of little interest or value.

Deconstructing the context in which unexpected observations arise is difficult, as it requires a capacity for closely following the phenomenology giving rise to something that may or may not be of any use. It is not only hard to know when pursuit of new avenues of investigation might be rewarded, but formulating the question to which the observation is an answer requires imagination and experience. Thus, Kuhn [71, p. 206] observes that a major value of quantified methods follows from the fact that numbers, sterile in themselves, “register departures from theory with an authority and finesse that no qualitative technique can duplicate.” Creating organizational environments capable of supporting full pivots in new directions is, of course, another matter entirely, but that is just what is entailed by the way science and society co-evolve [5, 54, 68, 69, 72, 77].

Continuing to accept summed ratings and multiparameter IRT models' undefined, unstable, uninterpretable, sample- and instrument-dependent unit quantities as necessary and unavoidable has proven itself as a highly effective means of arresting the development of psychology and the social sciences. The common practice of willfully mistaking ordinal ratings and IRT estimates for interval measures perpetuates the failure to even conceive the possibility that communities of research and practice could think and act together in the terms of common languages employed for their value as the media of communication and the rules by which exceptions are revealed. Continued persistence in this confusion has reached a quite perverse degree of pathological denial, given that the equivalence of measurement scales across the sciences was deemed “widely accepted” over 35 years ago [88, p. 169] but still has not fulfilled its potential in mainstream applications. Grimby et al. [62] are not unreasonable, from our point of view, in viewing the ongoing acceptance of ordinal scores and undefined numeric units in person-centered outcome measurement as a form of fraudulent malpractice.

The dominant paradigm's failure to distinguish numeric scores from measured quantities [10] commits the fundamental epistemological error of separating individual minds from the environmental context they inhabit [9, p. 493; 47]. Cognitive processes do not occur solely within brains, but must necessarily leverage scaffolded supports built into the external environment, such as alphabets, phonemes, grammars, dictionaries, printing presses, and quantitative unit standards' quality assurance protocols [65, 66, 75, 95, 106]. Metrological infrastructures define the processes by which real things and events in the world are connected with formal conceptual ideals and are brought into words with defined meanings, including common metrics' number words. And so, as Latour [72, pp. 249, 251] put it,

> Every time you hear about a successful application of a science, look for the progressive extension of a network. . . . Metrology is the name of this gigantic enterprise to make of the outside a world inside which facts and machines can survive.

Shared standards and common languages are the means by which we prepare minds on mass scales to better recognize and act on chance events. As Pasteur put it in 1854, “in the fields of observation, chance favors only the prepared mind” [40, p. 309]. Because currently popular measurement methods neither map the unfolding sequences of changes in health, performance, functionality, or learning, nor express differences in terms of defined unit quantities with stated uncertainties, nor reveal unexpected departures from theory, person-centered care lacks systematic ways of apprehending and communicating accidental and serendipitous events that might possess actionable value.

Identified models and metrological standards set up an alternative vision of broad ecosystems of interdependent and reproductively viable forms of social life. A key potential for productive innovations emerges here, since, as the populations of these forms of life grow, highly improbable combinations (mutations) become so frequent that their failure to occur is what becomes unlikely [41, p. 91]. In other words, multilevel metrologically-traceable systems of measurement create the combinations of construct theories' self-descriptive genotypes, instrument standard phenotypes, and mutable individual data needed for natural selection to amplify adaptively superior new forms of social life [91, 92] in a kind of epigenetic organism-environment integration. But if statistical descriptions of ordinal scores and IRT's varying unit estimates continue to be taken as satisfactory approaches to quantifying person-centered outcomes, it is only reasonable to expect continued perpetuation of the status quo vis-à-vis systematically and statistically concealed anomalies and exceptions to the rule that could otherwise lead in qualitatively productive new directions at both individual and aggregate levels.

1.4 Concluding Comments

Differences between centrally-planned data analytics and distributed metrological networks were a matter of concern for Ben Wright [123, 126, 127] not just in his steadfast focus on science over statistics but more broadly throughout his conception of measurement [46, 117]. In the last paragraph of his 1988 Inaugural Address to the AERA Rasch Measurement SIG, Wright ([124]; also see [125]) said:

> So, we come to my last words. The Rasch model is not a data model at all. You may use it with data, but it's not a data model. The Rasch model is a definition of measurement, a law of measurement. Indeed, it's *the* law of measurement. It's what we think we have when we have some numbers and use them as though they were measures. And it's the way numbers have to be in order to be analyzed statistically. The Rasch model is the condition that data must meet to qualify for our attention. It's our guide to data good enough to make measures from. And it's our criterion for whether the data with which we are working can be useful to us.

This recurring theme in Wright's work is also foregrounded on the first page of Wright and Masters' 1982 book [130]:

> Because we are born into a world full of well-established variables it can seem that they have always existed as part of an external reality which our ancestors have somehow discovered. But science is more than discovery. It is also an expanding and ever-changing network of practical inventions. Progress in science depends on the creation of new variables constructed out of imaginative selections and organizations of experience.

With his colleagues and students, Wright [122] advanced the ideas of item banking and adaptive item administration [8, 28, 29, 82, 129], individualized "kidmap" reports [26, 27, 131] and self-scoring forms [12, 50, 79, 80, 86, 126, 128, 132]. All of these depend on understanding measurement as operationalized via structurally invariant, anchored item estimates and networks of instruments read in a common language at the point of use [46, 117]. Wright's contributions to construct theorizing, instrument calibration, and individually-customized reports of special strengths and weaknesses span all three semiotic levels of complexity.

The chapters in this book build on and remain in dialogue with Wright's [127, p. 33] realization that "Science is impossible without an evolving network of stable measures." With increasing awareness of the metrological viability of instruments calibrated using Wright's ideas [25, 52, 85, 93], and the emergence of new consensus standards for uniform metrics [39], there is also increasing need for examples of the kind brought together in this book. We hope that the efforts begun by the contributors to this volume will inspire a growing sphere of imaginative and productive applications of person-centered outcome metrology.

References

1. J.R. Ackermann, *Data, Instruments, and Theory: A Dialectical Approach to Understanding Science* (Princeton University Press, 1985)
2. J. Aldrich, Autonomy. Oxf. Econ. Pap. **41**, 15–34 (1989)
3. D.D. Allen, M. Wilson, Introducing multidimensional item response modeling in health behavior and health education research. Health Educ. Res. **21**(suppl_1), 73–i84 (2006)
4. D. Andrich, Controversy and the Rasch model: A characteristic of incompatible paradigms? Med. Care **42**(1), I-7–I-16 (2004)
5. C. Audia, F. Berkhout, G. Owusu, Z. Quayyum, S. Agyei-Mensah, Loops and building blocks: A knowledge co-production framework for equitable urban health. J. Urban Health **98**(3), 394–403 (2021)
6. D. Bamber, J.P.H. van Santen, How many parameters can a model have and still be testable? J. Math. Psychol. **29**, 443–473 (1985)
7. E. Banks, The philosophical roots of Ernst Mach's economy of thought. Synthese **139**(1), 23–53 (2004)
8. M. Barney, W.P. Fisher Jr., Adaptive measurement and assessment. Annu. Rev. Organ. Psych. Organ. Behav. **3**, 469–490 (2016)
9. G. Bateson, *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology* (University of Chicago Press, 1972)
10. G. Bateson, Number is different from quantity. CoEvol. Q. **17**, 44–46 (1978) [Reprinted from pp. 53–58 in Bateson, G. (1979). *Mind and Nature: A Necessary Unity*. New York: E. P. Dutton]
11. D.M. Berwick, C.K. Cassel, The NAM and the quality of health care-inflecting a field. N. Engl. J. Med. **383**(6), 505–508 (2020)
12. W.R. Best, A Rasch model of the Crohn's Disease Activity Index (CDAI): Equivalent levels of ranked attribute and continuous variable scales, in *Crohn's Disease: Etiology, Pathogenesis and Interventions*, ed. by J. N. Cadwallader, (Nova Science Publishers, Inc, 2008), p. Chapter 5

13. A. Bilodeau, L. Potvin, Unpacking complexity in public health interventions with the Actor–Network Theory. Health Promot. Int. **33**(1), 173–181 (2018)
14. M. Black, *Models and Metaphors* (Cornell University Press, 1962/2019)
15. M. Black, More about metaphor, in *Metaphor and Thought*, ed. by A. Ortony, (Cambridge University Press, Cambridge, 1993), pp. 19–43
16. P. Black, M. Wilson, S. Yao, Road maps for learning: A guide to the navigation of learning progressions. Meas. Interdiscip. Res. Persp. **9**, 1–52 (2011)
17. A. Blok, I. Farias, C. Roberts (eds.), *The Routledge Companion to Actor-Network Theory* (Routledge, 2020)
18. R.K. Bode, A.W. Heinemann, P. Semik, Measurement properties of the Galveston Orientation and Amnesia Test (GOAT) and improvement patterns during inpatient rehabilitation. J. Head Trauma Rehabil. **15**(1), 637–655 (2000)
19. G. Bowker, S. Timmermans, A. E. Clarke, E. Balka (eds.), *Boundary Objects and beyond: Working with Leigh Star* (MIT Press, 2015)
20. J. Browne, S. Cano, S. Smith, Using patient-reported outcome measures to improve healthcare: Time for a new approach. Med. Care **55**, 901–904 (2017)
21. R. Bud, S.E. Cozzens (eds.), *SPIE Institutes: Vol. 9. Invisible connections: Instruments, Institutions, and Science*, ed. by R.F. Potter (SPIE Optical Engineering Press, 1992)
22. H. Burdick, A.J. Stenner, Theoretical prediction of test items. Rasch Meas. Trans. **10**(1), 475 (1996). http://www.rasch.org/rmt/rmt101b.htm
23. J.R. Busemeyer, Y.-M. Wang, Model comparisons and model selections based on generalization criterion methodology. J. Math. Psychol. **44**(1), 171–189 (2000)
24. H. Butterfield, *The Origins of Modern Science (Revised Edition)* (The Free Press, 1957)
25. S. Cano, L. Pendrill, J. Melin, W. Fisher, Towards consensus measurement standards for patient-centered outcomes. Measurement **141**, 62–69 (2019)
26. T.W. Chien, Y. Chang, K.S. Wen, Y.H. Uen, Using graphical representations to enhance the quality-of-care for colorectal cancer patients. Eur. J. Cancer Care **27**(1), e12591 (2018)
27. T.-W. Chien, W.-C. Wang, H.-Y. Wang, H.-J. Lin, Online assessment of patients' views on hospital performances using Rasch model's KIDMAP diagram. BMC Health Serv. Res. **9**, 135 (2009). https://doi.org/10.1186/1472-6963-9-135. or http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2727503/
28. B. Choppin, An item bank using sample-free calibration. Nature **219**, 870–872 (1968)
29. B. Choppin, Recent developments in item banking, in *Advances in Psychological and Educational Measurement*, ed. by D. N. M. DeGruitjer, L. J. van der Kamp, (Wiley, 1976), pp. 233–245
30. W. Cohen, L. Mundy, T. Ballard, A. Klassen, S. Cano, J.P. Browne, A. Pusic, The BREAST-Q in surgical research: A review of the literature 2009–2015. J. Plast. Reconstr. Surg. **69**, 149–162 (2016)
31. K. Cook, P.O. Monahan, C.A. McHorney, Delicate balance between theory and practice: Health status assessment and Item Response Theory. Med. Care **41**(5), 571–574 (2003)
32. T.A. Cook, *The Curves of Life* (Dover, 1914/1979)
33. A. Coulter, Measuring what matters to patients. Br. Med. J. **356**, j816 (2017)
34. L.H. Daltroy, M. Logigian, M.D. Iversen, M.H. Liang, Does musculoskeletal function deteriorate in a predictable sequence in the elderly? Arthritis Care Res. **5**, 146–150 (1992)
35. P. De Boeck, M. Wilson. Explanatory item response models: A generalized linear and nonlinearapproach. Statistics for Social and Behavioral Sciences). New York: Springer-Verlag (2004)
36. J. Dewey, in *Unmodern Philosophy and Modern Philosophy*, ed. by P. Deen, (Southern Illinois University Press, 2012)
37. S. Donetto, C. Chapman, S. Brearley, A.M. Rafferty, D. Allen, G. Robert, Exploring the impact of patient experience data in acute NHS hospital trusts in England: Using Actor-Network Theory to optimise organisational strategies and practices for improving patients' experiences of care. Health Serv. Deliv. Res. **14**, 156 (2019)

38. S.E. Embretson, Item Response Theory models and spurious interaction effects in factorial ANOVA designs. Appl. Psychol. Meas. **20**(3), 201–212 (1996)
39. EMPIR Project 18HLT04 NeuroMet, *Innovative Measurements for Improved Diagnosis and Management of Neurodegenerative Diseases*. https://www.lgcgroup.com/our-programmes/empir-neuromet/neuromet-landing-page/ (2022)
40. S. Finger, *Minds Behind the Brain: A History of the Pioneers and Their Discoveries* (Oxford University Press, 2004)
41. R.A. Fisher, Retrospect of the criticisms of the theory of natural selection, in *Evolution as a Process*, ed. by J. Huxley, A. C. Hardy, E. B. Ford, (George Allen & Unwin Ltd, 1954), pp. 84–98
42. W.P. Fisher Jr., The Rasch debate: Validity and revolution in educational measurement, in *Objective Measurement: Theory into Practice*, ed. by M. Wilson, vol. II, (Ablex Publishing Corporation, 1994), pp. 36–72
43. W.P. Fisher Jr., Objectivity in psychosocial measurement: What, why, how. J. Outcome Meas. **4**(2), 527–563 (2000)
44. W. P. Fisher, Jr. The standard model in the history of the natural sciences, econometrics, and the socialsciences. J. Phys. Conf. Ser. **238**(1) (2010). http://iopscience.iop.org/1742-6596/238/1/012016/pdf/1742-6596_238_1_012016.pdf.
45. W.P. Fisher Jr., Contextualizing sustainable development metric standards: Imagining new entrepreneurial possibilities. Sustainability **12**(9661), 1–22 (2020a)
46. W.P. Fisher Jr., Wright, Benjamin D. [Biographical entry], in *SAGE Research Methods Foundations*, ed. by P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, R. Williams, (Sage, 2020b). https://methods.sagepub.com/foundations/wright-benjamin-d
47. W.P. Fisher Jr., Bateson and Wright on number and quantity: How to not separate thinking from its relational context. Symmetry **13**, 1415 (2021a)
48. W.P. Fisher Jr., Separation theorems in econometrics and psychometrics: Rasch, Frisch, two fishers, and implications for measurement. J. Interdiscip. Econ., OnlineFirst, 1–32 (2021b)
49. W.P. Fisher Jr., *Metaphor and Measurement* (Submitted, in Review, 2022)
50. W.P. Fisher Jr., R.F. Harvey, K.M. Kilgore, New developments in functional assessment: Probabilistic models for gold standards. NeuroRehabilitation **5**(1), 3–25 (1995)
51. W.P. Fisher Jr., E.P.-T. Oon, S. Benson, Applying design thinking to systemic problems in educational assessment information management. J. Phys. Conf. Ser. **1044**, 012012 (2018). http://iopscience.iop.org/article/10.1088/1742-6596/1044/1/012012
52. W.P. Fisher Jr., A.J. Stenner, Theory-based metrological traceability in education: A reading measurement network. Measurement **92**, 489–496 (2016). http://www.sciencedirect.com/science/article/pii/S0263224116303281
53. W.P. Fisher Jr., A.J. Stenner, Ecologizing vs modernizing in measurement and metrology. J. Phys. Conf. Ser. **1044**, 012025 (2018). http://iopscience.iop.org/article/10.1088/1742-6596/1044/1/012025
54. W.P. Fisher Jr., M. Wilson, Building a productive trading zone in educational assessment research and practice. Pensamiento Educativo: Revista de Investigacion Educacional Latinoamericana **52**(2), 55–78 (2015)
55. G. Franck, The scientific economy of attention: A novel approach to the collective rationality of science. Scientometrics **55**(1), 3–26 (2002)
56. G. Franck, The economy of attention. J. Sociol. **55**(1), 8–19 (2019)
57. P. Galison, *Image and Logic: A Material Culture of Microphysics* (University of Chicago Press, 1997)
58. P. Galison, Trading zone: Coordinating action and belief, in *The Science Studies Reader*, ed. by M. Biagioli, (Routledge, 1999), pp. 137–160
59. J. Goldstein, M. Chun, D. Fletcher, J. Deremeik, R. Massof, Low vision research network study group. Visual ability of patients seeking outpatient low vision services in the United States. J. AMA Ophthalmol. **132**, 1169–1177 (2014)

60. J. Golinski, Is it time to forget science? Reflections on singular science and its history. Osiris **27**(1), 19–36 (2012)
61. C.V. Granger, R.T. Linn, Biologic patterns of disability. J. Outcome Meas. **4**(2), 595–615 (2000). http://jampress.org/JOM_V4N2.pdfs
62. G. Grimby, A. Tennant, L. Tesio, The use of raw scores from ordinal scales: Time to end malpractice? J. Rehabil. Med. **44**, 97–98 (2012)
63. L. Guttman, The basis for scalogram analysis, in *Measurement and Prediction*, Studies in Social Psychology in World War II. Volume 4, ed. by S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, J. A. Clausen, (Wiley, 1950), pp. 60–90
64. D.J. Haraway, Modest witness: Feminist diffractions in science studies, in *The Disunity of Science: Boundaries, Contexts, and Power*, ed. by P. Galison, D. J. Stump, (Stanford University Press, 1996), pp. 428–441
65. E. Hutchins, *Cognition in the Wild* (MIT Press, 1995)
66. E. Hutchins, The cultural ecosystem of human cognition. Philos. Psychol. **27**(1), 34–49 (2014)
67. D. Ihde, *Instrumental Realism: The Interface Between Philosophy of Science and Philosophy of Technology*, The Indiana Series in the Philosophy of Technology (Indiana University Press, 1991)
68. S. Jasanoff, *States of Knowledge: The Co-production of Science and Social Order*, International Library of Sociology (Routledge, 2004)
69. S. Jasanoff, The practices of objectivity in regulatory science, in *Social Knowledge in the Making*, ed. by C. Camic, N. Gross, M. Lamont, (University of Chicago Press, 2011), pp. 307–338
70. T.C. Koopmans, O. Reiersøl, The identification of structural characteristics. Ann. Math. Stat. **XXI**, 165–181 (1950)
71. T.S. Kuhn, The function of measurement in modern physical science. Isis **52**(168), 161–193 (1961/1977). (Rpt. in T. S. Kuhn, (Ed.). (1977). The essential tension: Selected studies in scientific tradition and change (pp. 178–224). University of Chicago Press)
72. B. Latour, *Science in Action: How to Follow Scientists and Engineers Through Society* (Harvard University Press, 1987)
73. B. Latour, Postmodern? No, simply amodern: Steps towards an anthropology of science. Stud. Hist. Phil. Sci. **21**(1), 145–171 (1990)
74. B. Latour, *We Have Never Been Modern* (Harvard University Press, 1993)
75. B. Latour, Cogito ergo sumus! Or psychology swept inside out by the fresh air of the upper deck: Review of Hutchins' *Cognition in the Wild*, MIT Press, 1995. Mind Cult. Activity Int. J. **3**(192), 54–63 (1995)
76. B. Latour, To modernise or ecologise? That is the question, in *Remaking Reality: Nature at the Millennium*, ed. by B. Braun, N. Castree, (Routledge, 1998), pp. 221–242
77. B. Latour, *Reassembling the Social: An Introduction to Actor-Network-Theory*, Clarendon Lectures in Management Studies (Oxford University Press, 2005)
78. J.M. Linacre, Stochastic Guttman order. Rasch Meas. Trans. **5**(4), 189 (1991). http://www.rasch.org/rmt/rmt54p.htm
79. J.M. Linacre, Instantaneous measurement and diagnosis. Phys. Med. Rehabil. State Art Rev. **11**(2), 315–324 (1997). http://www.rasch.org/memo60.htm
80. J. Liu, Development and translation of measurement findings for the motivation assessment for team readiness, integration, and collaboration self-scoring form. Am. J. Occup. Ther. **72**-(4_Supplement_1), 7211500015p1 (2018)
81. J. Lumsden, Tests are perfectly reliable. Br. J. Math. Stat. Psychol. **31**, 19–26 (1978)
82. M.E. Lunz, B.A. Bergstrom, R.C. Gershon, Computer adaptive testing. Probabilistic Conjoint Measurement. A Special Issue of the Int. J. Educ. Res. (W.P. Fisher, Jr., B.D. Wright, eds.), **21**(6), 623–634 (1994)
83. E. Mach, *The Science of Mechanics: A Critical and Historical Account Of Its Development*, Trans. T.J. McCormack, 4th ed. (The Open Court Publishing Co., 1883/1919)

84. L. Mari, M. Wilson, An introduction to the Rasch measurement approach for metrologists. Measurement **51**, 315–327 (2014)
85. L. Mari, M. Wilson, A. Maul. *Measurement Across the Sciences*, R. Morawski, G. Rossi, others, eds., Springer Series in Measurement Science and Technology (Springer, 2021)
86. G.N.Masters, R.J. Adams, J. Lokan, Mapping student achievement. Probabilistic Conjoint Measurement, A Special Issue of the Int. J. Educ. Res. (W.P. Fisher, Jr., B.D. Wright, eds.), **21**(6), 595–610 (1994)
87. J. Melin, S. Cano, L. Pendrill, The role of entropy in construct specification equations (CSE) to improve the validity of memory tests. Entropy **23**(2), 212 (2021)
88. L. Narens, R.D. Luce, Measurement: The theory of numerical assignments. Psychol. Bull. **99**(2), 166–180 (1986)
89. N.J. Nersessian, Maxwell and "the method of physical analogy": Model-based reasoning, generic abstraction, and conceptual change, in *Reading Natural Philosophy: Essays in the History and Philosophy of Science and Mathematics*, ed. by D. Malament, (Open Court, 2002), pp. 129–166
90. J. O'Connell, Metrology: The creation of universality by the circulation of particulars. Soc. Stud. Sci. **23**, 129–173 (1993)
91. H.H. Pattee, Universal principles of measurement and language functions in evolving systems, in *Complexity, Language, and Life: Mathematical Approaches*, ed. by J. L. Casti, A. Karlqvist, (Springer, 1985), pp. 268–281
92. H.H. Pattee, J. Raczaszek-Leonardi, *Biosemiotics. Vol. 7: Laws, Language and Life: Howard Pattee's Classic Papers on the Physics of Symbols with Contemporary Commentary*, M. Barbieri, J. Hoffmeyer, eds., (Springer, 2012)
93. L. Pendrill, *Quality Assured Measurement: Unification Across Social and Physical Sciences*, R. Morawski, G. Rossi, others, eds., Springer Series in Measurement Science and Technology (Springer, 2019)
94. L. Pendrill, W.P. Fisher Jr., Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. Measurement **71**, 46–55 (2015)
95. E. Petracca, S. Gallagher, Economic cognitive institutions. J. Inst. Econ. **16**(6), 747–765 (2020)
96. G. Rasch, *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, 1960)
97. R.M. Roberts, *Serendipity: Accidental Discoveries in Science* (Wiley, 1989)
98. E. San Martin, J. Gonzalez, F. Tuerlinckx, Identified parameters, parameters of interest, and their relationships. Meas. Interdiscip. Res. Persp. **7**(2), 97–105 (2009)
99. E. San Martin, J. Gonzalez, F. Tuerlinckx, On the unidentifiability of the fixed-effects 3 PL model. Psychometrika **80**(2), 450–467 (2015)
100. E. San Martin, J.M. Rolin, Identification of parametric Rasch-type models. J. Stat. Plan. Inference **143**(1), 116–130 (2013)
101. S.L. Star, The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving, in *Proceedings of the 8th AAAI Workshop on Distributed Artificial Intelligence, Technical Report*, (Department of Computer Science, University of Southern California, 1988/2015). (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)
102. S.L. Star, J.R. Griesemer, Institutional ecology, 'translations,' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. Soc. Stud. Sci. **19**(3), 387–420 (1989)
103. S.L. Star, K. Ruhleder, Steps toward an ecology of infrastructure: Design and access for large information spaces. Inf. Syst. Res. **7**(1), 111–134 (1996)
104. A.J. Stenner, W.P. Fisher Jr., M.H. Stone, D.S. Burdick, Causal Rasch models. Front. Psychol. Quant. Psychol. Meas. **4**(536), 1–14 (2013)

105. M.H. Stone, Substantive scale construction. J. Appl. Meas. **4**(3), 282–297 (2003)
106. J. Sutton, C.B. Harris, P.G. Keil, A.J. Barnier, The psychology of memory, extended cognition, and socially distributed remembering. Phenomenol. Cogn. Sci. **9**(4), 521–560 (2010)
107. UK Department of Health, UK PROMS Programme (2022), https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/patient-reported-outcome-measures-proms
108. US Food and Drug Administration, FDA Patient-Focused Drug Development Guidance Series for Enhancing the Incorporation of the Patient's Voice in Medical Product Development and Regulatory Decision Making (2022), https://www.fda.gov/drugs/development-approval-process-drugs/fda-patient-focused-drug-development-guidance-series-enhancing-incorporation-patients-voice-medical
109. A. van Helden, T. L. Hankins (eds.), *Instruments (Vol. 9). Osiris: A Research Journal Devoted to the History of Science and Its Cultural Influences* (University of Chicago Press, 1994)
110. N.D. Verhelst, C.A.W. Glas, The one parameter logistic model, in *Rasch Models: Foundations Recent Developments, and Applications*, ed. by G. H. Fischer, I. W. Molenaar, (Springer, 1995), pp. 215–237
111. M. Wilson (ed.), *National Society for the Study of Education Yearbooks*, Vol. 103, Part II: Towards Coherence Between Classroom Assessment and Accountability (University of Chicago Press, 2004)
112. M.R. Wilson, *Constructing Measures: An Item Response Modeling Approach* (Lawrence Erlbaum Associates, 2005a)
113. M. Wilson, Subscales and summary scales: Issues, in *Outcomes Assessment in Cancer: Measures, Methods and Applications*, ed. by J. Lipscomb, C. C. Gotay, C. Snyder, (Cambridge University Press, 2005b), pp. 465–479
114. M. Wilson, Cognitive diagnosis using item response models. Zeitschrift Für Psychologie/J. Psychol. (Special Issue: Current Issues in Competence Modeling and Assessment) **216**(2), 74–88 (2008)
115. M. Wilson, Making measurement important for education: The crucial role of classroom assessment. Educ. Meas. Issues Pract. **37**(1), 5–20 (2018)
116. M. Wilson, W.P. Fisher Jr., Preface: 2016 IMEKO TC1-TC7-TC13 Joint Symposium: Metrology Across the Sciences: Wishful Thinking? J. Phys. Conf. Ser. **772**(1), 011001 (2016)
117. M. Wilson, W. P. Fisher Jr. (eds.), *Psychological and Social Measurement: The Career and Contributions of Benjamin D. Wright*, ed. by M. G. Cain, G. B. Rossi, J. Tesai, M. van Veghel, K.-Y. Jhang, Springer Series in Measurement Science and Technology (Springer, 2017). https://link.springer.com/book/10.1007/978-3-319-67304-2
118. M. Wilson, W.P. Fisher Jr., Preface of special issue, Psychometric Metrology. Measurement **145**, 190 (2019)
119. M. Wilson, P. Gochyyev, Having your cake and eating it too: Multiple dimensions and a composite. Measurement **151**, 107247 (2020)
120. M.N. Wise, Precision: Agent of unity and product of agreement. Part III—"Today precision must be commonplace", in *The Values of Precision*, ed. by M. N. Wise, (Princeton University Press, 1995), pp. 352–361
121. A.W. Woolley, E. Fuchs, Collective intelligence in the organization of science. Organ. Sci. **22**(5), 1359–1367 (2011)
122. B.D. Wright, Solving measurement problems with the Rasch model. J. Educ. Meas. **14**(2), 97–116 (1977)
123. B.D. Wright, Despair and hope for educational measurement. Contemp. Educ. Technol. **3**(1), 281–288 (1984)
124. B.D. Wright, 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 Meas. Trans. **2**, 25–32 (1988a). http://www.rasch.org/rmt/rmt23.htm. (Rpt. in J. M. Linacre, (Ed.). (1995). Rasch Measurement Transactions, Part 1 (pp. 25–32). MESA Press)
125. B.D. Wright, Useful measurement through concurrent equating and one-step (concurrent) item banking. Rasch Meas. Trans. **2**(2), 24 (1988b). http://www.rasch.org/rmt/rmt22f.htm

126. B.D. Wright, Fundamental measurement for outcome evaluation. Phys. Med. Rehabil. State Art Rev. **11**(2), 261–288 (1997a)

127. B.D. Wright, A history of social science measurement. Educ. Meas. Issues Pract. **16**(4), 33-45–33-52 (1997b)

128. B.D. Wright, Benjamin D. Wright's annotated KeyMath diagnostic profile. Rasch Meas. Trans. **25**(4), 1350 (2012). https://www.rasch.org/rmt/rmt254.pdf

129. B.D. Wright, S.R. Bell, Item banks: What, why, how. J. Educ. Meas. **21**(4), 331–345 (1984). http://www.rasch.org/memo43.htm

130. B.D. Wright, G.N. Masters, *Rating Scale Analysis: Rasch Measurement* (MESA Press, 1982)

131. B.D. Wright, R.J. Mead, L.H. Ludlow, *KIDMAP: Person-by-Item Interaction Mapping*, MESA Memorandum #29 (MESA Press, Chicago, 1980). http://www.rasch.org/memo29.pdf

132. B.D. Wright, M.H. Stone, *Best Test Design: Rasch Measurement* (MESA Press, 1979)

Chapter 2
A Clinician's Guide to Performance Outcome Measurements

Anna G. Mayhew

Abstract The role of the clinical evaluator in delivering measurements of motor performance is a key one. Good assessment enables effective management and in the trial setting enables partners to be confident in the reliability of their motor endpoints as a true reflection of the compound's efficacy. Here we explore how Rasch analysis enhances more traditional evaluations of reliability and validity in the evaluation of performance using practical examples to explain why and discuss the importance of the role of clinicians in the interpretation and decision-making process when using Rasch analysis. The North Star Ambulatory Assessment designed for use in ambulant Duchenne muscular dystrophy illustrates the role of these methods. The role of the clinician is to ensure that wherever we assess a patient—in clinic, at home or via video link—the measured value and clinical meaning can be understood and potentially equated in relationship to other similar scales conducted in an alternative forum. Our ability to interact with and be aware of the individual at the heart of all these assessments means we are uniquely placed to ensure best care is given.

Keywords Clinician perspective · Ambulatory assessment · Duchenne muscular dystrophy · Measurement applications

2.1 Introduction

For clinicians, and especially for therapists who measure client performance, our role in delivering measurements effectively and consistently has never been more critical. This is in part due to the success of clinical trials in rare diseases which affect motor performance, and which will therefore often involve measurements of motor

A. G. Mayhew (✉)
The John Walton Muscular Dystrophy Research Centre, Translational and Clinical Research Institute, Newcastle upon Tyne, UK

Newcastle University and Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK
e-mail: anna.mayhew@newcastle.ac.uk

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3_2

performance as a primary outcome. Ultimately the success of the trial – correctly identifying the therapy as effective or not – rests on the ability of this endpoint to detect an actual improvement correctly and not record a negative trial due to the scale's inadequacies as a measurement tool. This has required us as clinicians to raise our game in terms of conducting performance outcome assessments consistently across a large number of centers, often in different countries [14]. It has also, of course, shone a spotlight on the validity of the chosen outcome instruments to measure what they say they will.

These requirements should not distract us from our additional role of measuring performance and function in order to evaluate the effect of a given non-medicinal intervention or therapy such as exercise or orthotics. A physiotherapist evaluating a child using a motor scale in a clinic may use the information gained to evaluate disease progression as well as guide the family in the provision of equipment and predicting the need for additional support in the future. This role is not separate from the physiotherapist's role in a clinical trial where the motor scale is a primary outcome and may also be essential to ensure that the necessary standard of care is provided to the participant. A good example is in Duchenne muscular dystrophy (DMD) where the North Star Ambulatory Assessment (NSAA) can identify the need for a change in steroid management or identify a loss of range of movement in the ankle joints requiring a program of stretches and orthotics as a recommended approach in the published standard of care guidelines [2].

As clinicians we have an additional responsibility. Person centred outcomes require the human touch to deliver these appropriately. Asking patients to repeat the same measure time and again over a long period requires sensitivity as well as a consistent delivery. Care cannot be just about measurement; it is about appreciating the difficulties experienced by patients who find themselves getting worse and finding tasks harder at each visit. Our focus is to remember our role in using data to not just drive a natural history study or a clinical trial but also in attending to how the information can be applied to management, and to how our words and actions as we ask patients to do yet more things can make a difference between them feeling just measured or ultimately cared for. This is of course another good reason to always involve patients in the evaluation and development of outcome measures and is particularly important in patient reported outcome measures. Their early and direct involvement will ensure the necessary sensitivities and clarity.

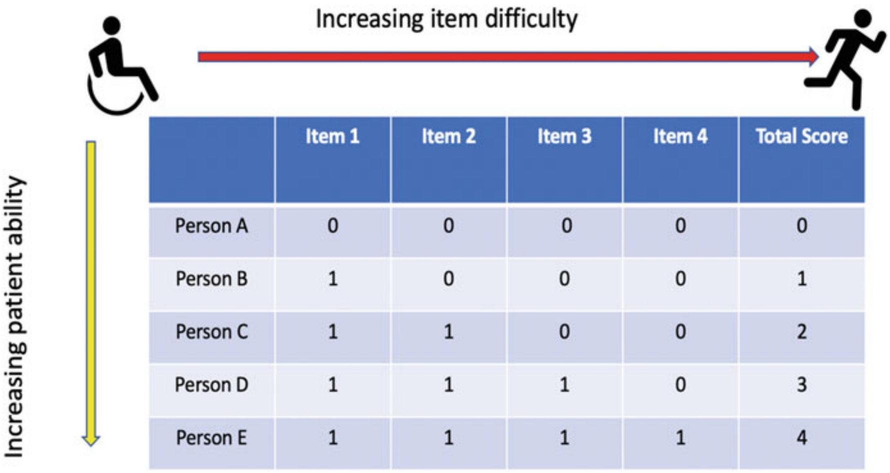
With this in mind this chapter aims first to illuminate the specific benefits of Rasch analysis in the evaluation of performance, using practical examples to explain why these methods enhance more traditional evaluations of reliability and validity. Second I will discuss the importance of the role clinicians play in the interpretation and decision making process when using Rasch analysis. Third, I explore the role of the clinician alongside other key stakeholders to ensure that wherever we assess a patient—in clinic, at home, or via video link—the value and clinical meaning of measurements can be understood and potentially equated in relationship to other similar scales conducting in an alternative forum. This speaks especially to the role of measurement in telemedicine.

2.2 Rasch Analysis Explained and Why Clinician Involvement Is Key

In this section we are going to use various real-world scales to illustrate measures of scale robustness utilized in modern psychometric methods and importantly highlight the complementary role of the clinician in interpretation of the analysis. The roles we as clinicians play in making ongoing decisions about a scale's development—decisions that enhance the analysis and benefit the scale's evolution—are stories often left untold within our medical publications. These accounts convey key components of the development of patient centred outcome measures requiring multi-disciplinary collaborations with patients and families [15].

2.2.1 *Guttman Principle*

Rasch methodology compares our real-world data (ordinal in nature) with the perfect mathematical scale (interval level data), where each item in the scale contributes to the measurement of one thing. This single thing is often termed a construct or a concept. In this perfect scale every single item contributes to the measurement of this one construct. It is also essential that the scoring options for each item incrementally improve or increase in the same way as the total score increases. Additionally, in this cumulative scale, the difficulty of each item is always ranked in the same order – there is a hierarchy of difficulty. This perfect scale is best summed up as the Guttman distribution and it is important for clinicians to understand this model to understand their own data. Figure 2.1 illustrates the perfect scale where each item is scored as a


Increasing item difficulty
Increasing patient ability

	Item 1	Item 2	Item 3	Item 4	Total Score
Person A	0	0	0	0	0
Person B	1	0	0	0	1
Person C	1	1	0	0	2
Person D	1	1	1	0	3
Person E	1	1	1	1	4

Fig. 2.1 Illustration of the Guttman principle

0 (unable) or 1 (able). What makes this such a powerful scale is that as the item difficulty does not vary, so knowing the total score means you know exactly how an individual scored each item. For example, if the total score was 3 out of top score of 4, you know the individual was able (score 1) to do item 1, 2 and 3 but not able to do item 4.

So, let's use an example and see how the North Star Ambulatory Assessment (NSAA) which is used to measure motor performance in boys and young men who can walk and who have a diagnosis of DMD [18] compares to the perfect mathematical scale.

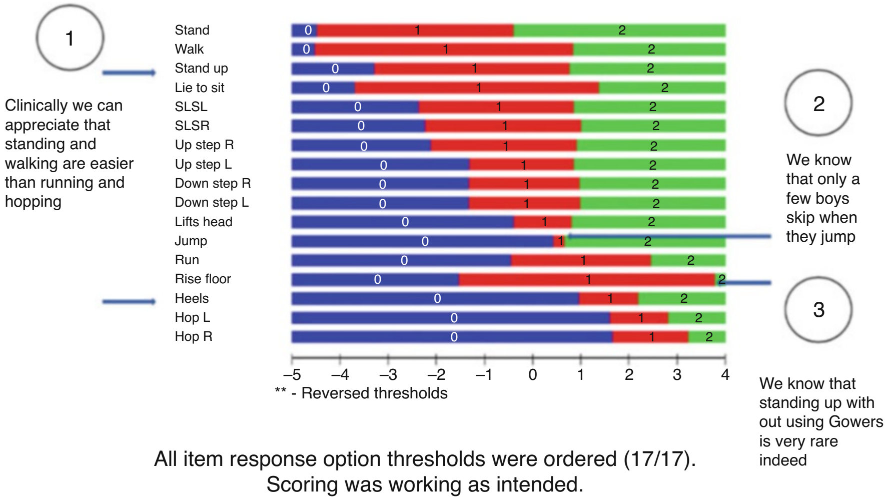
2.2.2 *The North Star Ambulatory Assessment*

The NSAA is a multiple item rating scale (17 items) with three ordered response categories (2, 1, or 0) which are summed to give a total score. Items are scored either 2 ('normal' with no obvious modification of activity), 1 (modified method but achieves goal independent of physical assistance from another), or 0 (unable to achieve independently). A total 'ambulatory function' score is generated by summing the ratings across the items. A higher score indicates better motor performance. The analyses presented here have been previously [9, 11] published but we will add clinical subtext, especially around the role of the physiotherapist in making decisions about the scale's content.

The first test which we conducted, which was designed to reassure clinicians, showed how the RUMM2030 program [1]- a software package used for measurement scaling data analyses – ranked the 17 items in order of difficulty compared to the expert opinion of five neuromuscular physiotherapists who regularly used the NSAA. This enabled a comparison to be made between the Rasch item hierarchy and clinical expectation. Consistency was examined qualitatively and statistically (Spearman's rho). The clinical utility of the total score was validated, as both clinicians and Rasch analysis reported stand and walk as easiest, while jump and hop were hardest; consequently, the Spearman rho was high.

Next, a key test of a scale is the ordering of the response categories. Each NSAA item has multiple response categories which reflect an ordered continuum of better ambulatory function (2, 1, and 0). The point at which performance on an item moves from one score to another is called a threshold. Although this ordering may appear clinically sensible at the item level, it must also work when the items are combined to form a set. Rasch analysis tests this statistically and graphically. When the response options are working as expected, this provides some important evidence for the validity of the scale.

This matters as each item's scoring options must rank numerically higher as a higher total score is achieved. If they do not the total score can be questioned. This lack of incremental improvement or decline is termed disordered thresholds and can occur for many reasons. Perhaps the wording or interpretation of an items scoring options is not clear or too complicated (too many response items). Perhaps the



Clinically we can appreciate that standing and walking are easier than running and hopping

We know that only a few boys skip when they jump

We know that standing up with out using Gowers is very rare indeed

All item response option thresholds were ordered (17/17). Scoring was working as intended.

Fig. 2.2 Threshold Map for NSAA Items in ranked order of difficulty according to Rasch analysis. (“Gowers” = adapted method of getting up from floor when muscles are weak)

scoring options are different strategies rather than hierarchical changes. Or perhaps the patient sample does not include people performing at that level.

For example, in the NSAA, item 3 describes the ability to “stand up” from a chair. A score of 2 means the patient is strong and can stand up without using their arms or moving their feet. A score of 1 is defined as uses an adaptation to the start position in order to stand up. This could be achieved by widening the feet or using their arms. However, if the scale attempted to define “uses arms” as always meaning a person doing so was stronger than one who moved their legs, the thresholds maybe be disordered because some may use one strategy and others, another, even though their weaknesses may be similar.

In Fig. 2.2 we can look at the map of the response categories and see how we can interpret them using our clinical knowledge. If this map did not make sense, this would suggest we need to relook at our scale and take into consideration some of the points raised above in the example of “stand up.” For the NSAA, “stand up” has ordered thresholds suggesting that the three scoring options were working as planned and the grouping of strategies into a score of 1 was clinically sensible and unambiguous. It is this clinical knowledge of a how a condition progresses, and how boys and young men present, that means as clinicians we lie at the heart of making changes to a scale. A scaling analyst may suggest we need to review the scoring options, but it would be clinicians who made changes using expert knowledge on a disease and its progression.

Next, we wanted to examine the match between the range of ambulatory function measured by the NSAA and the range of ambulation measured in the sample of children. Figure 2.3 shows the adequate targeting between the distribution of person measurements (upper histogram) and the distribution of item locations (lower

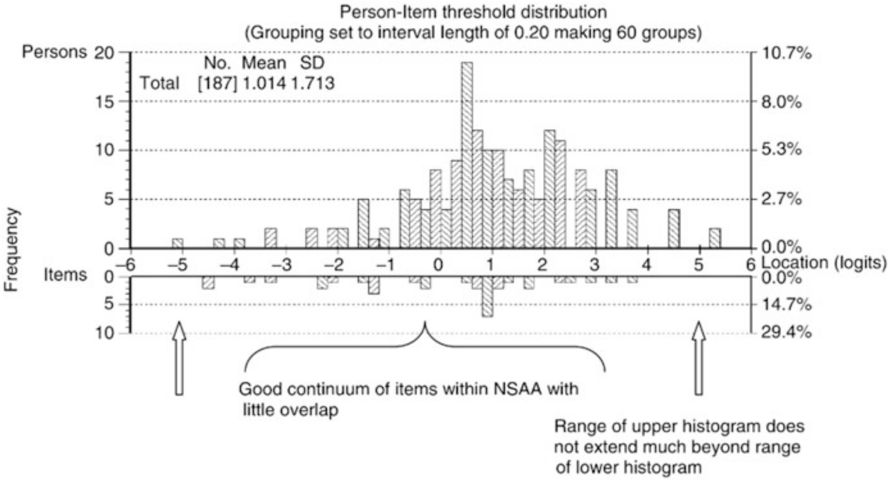


Fig. 2.3 Person-item location distribution

histogram). This analysis informs us as to how suitable the sample is for evaluating the NSAA and how suitable it is for measuring the sample. This is often called targeting and better targeting means the scale is working well independent of the sample, as well as working for the sample tested.

Traditionally, we understand this as a floor and ceiling effect; Rasch analysis allows us to also examine the spread of items. This allows us to ask questions such as "Are there some children we are not measuring well?" (gaps in the scale) or are there some levels we are over assessing (bunching of items)? For the NSAA there is small ceiling and as the disease is progressive there will be a floor as boys and young men lose ambulation. Clinicians will immediately recognize this as appropriate for a sample and can use their clinical knowledge to suggest additional items to bridge gaps in the current scale, to reduce any ceiling or floor effects, or suggest items to remove, as they measure the same level of ability. For instance, the cluster of items at about 1 logit on the horizontal scale relates to the box step items, some of which may justifiably be removed from a scale that aims to measure level of ability only.

Next, we examined the items of the NSAA to see if they worked together (fit) cohesively. If they are measuring more than one thing, some items would be inappropriate to the overall interpretation of the ratings and to considering the total score as a basis for the measurement of ambulatory motor function. When items do not work together (misfit) in this way, the validity of a scale is questioned. The methods for examining this using Rasch methods are four-fold and it is best to interpret all four tests together in the context of your clinical experience, including:

1. Fit residuals (log residuals – which should be between the range of + − 2.5, depending on sample size and test length);
2. x^2 values (item–trait interaction) – which need to be numerically similar to each other;

3. t-test for unidimensionality – a standard test of whether the scale measures one thing or not; and
4. item characteristic curves – which tell you about whether the way an item works in real life is as the model predicts.

For this measure of fit let us use the NSAA again as an example of how a clinician's experience can assist interpretation and steps taken. For fit residuals three items misfit but two of these only slightly (climb and second box step right leg first). One item however showed significant misfit: lifts head in supine position. This inconsistency makes clinical sense. It is a motor task impacted by DMD, and so it was included in the original scale because performance often improved when steroids were started. That is, it was originally included even though it does not directly relate to ambulatory function. As the inconsistency of the ratings for the "lifts head" item makes clinical sense and this result was reflected in other measures of fit, such as the item characteristic curve (the item did not change in the same way as other items did as the disease progressed), and the x^2 values (high compared to other values), it was decided that in the linearized scale (where the ordinal level data is converted to interval level measurement) that "lifts head" would be removed from the total score prior to transformation. This decision was also approved by clinicians as clinically sensible.

Another analytical tool is the Person Separation Index (PSI), a reliability statistic comparable to Cronbach's alpha. Higher values indicate greater reliability. The NSAA had a high reliability which confirmed findings from a study that examined reliability using traditional methods (intra-class correlations) [13].

When dependency was examined for the NSAA several pairs of items were found to be dependent. This occurs when the score on one item directly influences the score on another item. If this happens, measurement estimates can be biased and reliability (PSI) is artificially elevated. This was true for the NSAA items measuring right and left sides (stand on one leg, hop, climb and descend box step). The reason the scale includes both sides is that, though DMD is not predominantly an asymmetrical disorder, differences can influence functioning at home and in the community and may require individual management – a right ankle splint for instance. Given the clinical importance of measuring both sides the decision was made to remove the scores from one side of the body to see if this influenced the reliability. As their removal did not change PSI value significantly clinicians decided to keep the bilateral measures given their clinical utility in assessing for standards of care.

Next, we assessed stability of the scale (differential item functioning) to understand if different subgroups performed items in a similar way regardless of other differences. For instance, for one group we assessed different treatment regimes of steroids, but for other cohorts we wanted to see if gender or age made a difference to scoring stability. For the NSAA, the type of regime did not influence the ways boys scored the items which provides reassurance that it can be used in all ambulant individuals with DMD.

Finally, we examined how closely the summed NSAA scores, which are by definition ordinal, correspond to the interval-level measurements. Basically, the question is, how close is the real-world data to looking like a ruler? Rasch analysis

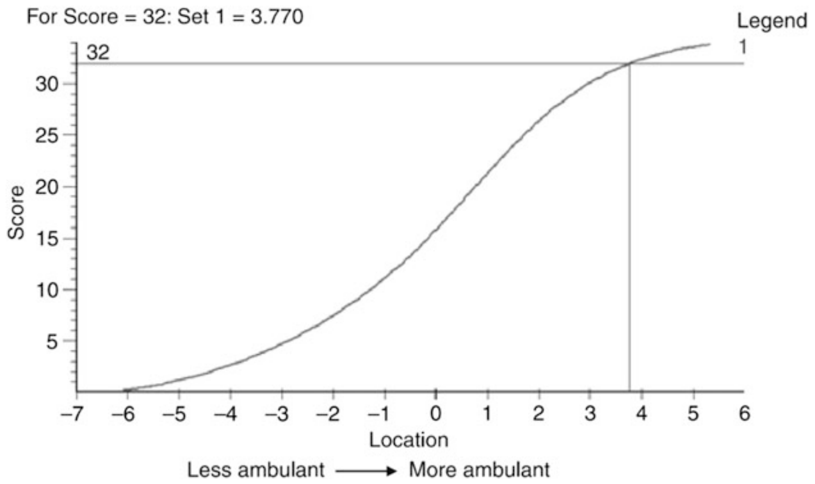


Fig. 2.4 Ordinal score to interval measure transformation graph

estimates linear measurements from raw scores, a relationship that can be illustrated in a graphical plot known as a logistic ogive (Fig. 2.4). This figure shows that the change in interval measurements associated with a one-point change in the NSAA total score is nonlinear: it varies across the range of the scale, just as ordinal scores always do.

Tests of ordered scoring options, fit, stability, dependency and reliability support the scaling of the NSAA as an interval level measuring instrument. This means that change is measured and defined consistently across the scale. Therefore, regardless of how strong or weak an individual was, the change measurement means the same thing.

This NSAA "ruler" was then used to examine the differences in response to two steroid treatment regimes, which, in the context of having identified differences that make a difference, could then be used to estimate a minimally important difference. Key to our clinical interpretation was describing minimally important differences (MIDs) in terms of meaningful change to the individuals and their families. The proposed MIDs could be equated to significant 'milestones' of loss. In more able males, a fall in interval level measurements from 90 to 80 (raw score 31–29) means they can no longer hop, and a fall from 50 to 40 (raw score 16–11) fits with an inability to rise independently from the floor. In weaker males, a fall from 21 to 11 (raw score 3–1) means they lose the ability to stand still. This also fits with our clinical understanding of the hierarchy of difficulty of items which was reported earlier.

Subsequent publications using this linearized scale have gone on to report change and average rates of decline [17], and have used the total score to identify different clusters within the wider cohort showing different patterns of change over time and the degree to which these patterns may be associated with age [16].

Finally, Rasch analysis allows comparisons of one scale with another that purports to measure the same construct. A good published example of this kind of equating is the development of the Revised Upper Limb Scale (RULM) [12]. Here an existing scale designed for weak young patients with spinal muscular atrophy (SMA) was adapted (by clinicians and physiotherapists assisted by individuals with SMA) to measure stronger individuals. The RULM has proven useful in this population, both in the clinic and in clinical trials [4, 21].

Work is underway to further test the RULM in other populations, and to use it to advance our understanding of the relationship between PROMs and functional scales [10]. Another example in this vein concerns a study conducted in the clinically important context of dysferlinopathy—a type of limb girdle muscular dystrophy—that involved the equating of two scales (the NSAA and the Motor Function Measure 20) brought together to produce a novel scale suitable for ambulant and non-ambulant individuals. The resulting North Star Assessment for Limb Girdle Type Muscular Dystrophies (NSAD) [7] is now being validated in a study incorporating a larger group of muscular dystrophies.

Figure 2.5 illustrates how the RULM measured more able individuals and created distinctions among those who were clustered at the ceiling of the original ULM. Advanced measurement modelling methods provide valuable insights into the suitability of any scale's ability to measure change effectively. Clinicians can benefit from this and should partner with measurement colleagues to build better scales.

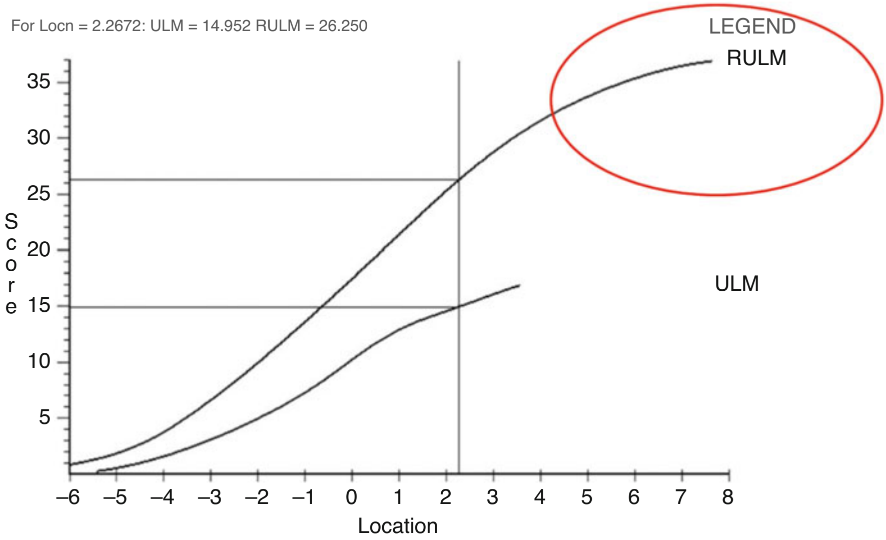


Fig. 2.5 Equating of RULM and ULM. The vertical axis shows that an ordinal score on the ULM of 15 is equivalent to a 26 on the RULM. The entry item was excluded from the total score on the RULM

2.3 Top Tips When Choosing Which Scale of Performance to Use

It is important that any scale that you use as a measurement tool within your clinical setting suits the patient group in question. You will want a scale sensitive to clinically significant change, perhaps in more than one direction. Because we are busy, clinicians often think that using an existing scale and applying it to a new population is the best and quickest way forward, and that any problematic issues can be sorted out at a later stage. Then, after we have used a convenient scale for some time, we may be lulled into accepting it despite various shortcomings. Later, after accruing quite a lot of longitudinal data, even though it is not a great scale, we don't want to change because we would then lose comparable continuity with our old data!

So where can we go from here? Historical data offers powerful opportunities for evaluating the internal workings of a scale. This can then guide the experts to adapt it to improve its measurement quality or replace it with an alternative scale which others have developed. It may be that these options are not relevant to your situation, and you need to start from scratch. In that case, we suggest you make good use of the literature. The rather lengthy iterative processes of patient involvement, scale development, testing, changes and re-evaluation [3, 5, 15] although a rewarding process, comprises "a large set of wheels that not everyone has time to turn".

Beware of scales designed for conditions and populations different from those of concern to you. Ask tough questions of the scales you intend to use or have been asked to use. Does it measure what it says it measures? How appropriate is an infant scale for use in adults? How does a scale designed for upper motor neuron problems work in a disease where the main issue is fatigue? These questions can only be answered if you know your disease, how it progresses, and with a team of experts who are critically engaged—and these experts must include patients and their families.

Don't be afraid of research literature that uses a lot of statistical tests. Learn with others, ask those who know more about measurement for advice, and always hold up a scale to your clinical sensibility for careful examination. A measure of performance that is reported to be highly reliable may not be valid. Conversely, a scale with valid content may not be precise enough to make distinctions at the degree of clinical significance needed to support your decision process. You may be acutely aware that a scale of motor performance does not tell you what you need to know about your weakest patients (so find an alternative). Or, from a practical point of view, you may know that the activities involved are not safe in a particular group. Essentially, ask questions!

2.4 The Future: Telemedicine?

Global reaching circumstances mean that as clinicians we have been seeing our patients and families at more than arm's length. Our current scales performed face to face have not been possible over the phone or at the very least have been difficult to perform via video conferencing methods. We have an opportunity here to compare our face-to-face assessments with those done remotely by video or we can adapt our current measures to be more suitable for use in the home. This work is already underway with some success in demonstrating the value of these remote measures [8]. We can see how our current scales relate specifically to an alternative model such as a patient reported outcome that measures the same construct. The issues and comparability of these different measures of performance can be evaluated using some of the techniques touched upon in this chapter (equating rating scales) and are described in more detail elsewhere [6]. This method of comparing scales that purport to measure the same construct is supported by regulatory authorities, often described as triangulation and can be seen in action using advanced measurement modelling in diseases such as dementia [19]. We may also find a new world engages us more with digital health technologies and it will be key that these novel methods summarize this sensor data into meaningful outcome measures [20] and we have a role to ensure that any novel measure has purpose and meaning.

As clinicians we are highly privileged to be directly involved with patients. The future may mean we need to take into better account the large amount of their lives that they conduct when they are not in clinic with us. Our ability to listen, interpret and support individuals will play a significant role and we must ensure we are part of future scale development and must not be afraid to embrace numbers because they only matter when we attach meaning to them that matters to individuals.

References

1. D. Andrich, B. Sheriden, G. Luo, RUMM 2030: Rasch unidimensional models for measurement. Crawley, Western Australia: University of Western Australia (2017)
2. D.J. Birnkrant, K. Bushby, C.M. Bann, S.D. Apkon, A. Blackwell, D. Brumbaugh, L.E. Case, P.R. Clemens, S. Hadjiyannakis, S. Pandya, N. Street, J. Tomezsko, K.R. Wagner, L.M. Ward, D.R. Weber, Diagnosis and management of Duchenne muscular dystrophy, Part 1: Diagnosis, and neuromuscular, rehabilitation, endocrine, and gastrointestinal and nutritional management. Lancet Neurol. **17**(3), 211–212 (2018)
3. S.J. Cano, L.R. Pendrill, J. Melin, W.P. Fisher, Towards consensus measurement standards for patient-centered outcomes. Meas. J. Int. Meas. Confed. **141**, 62–69 (2019)
4. G.B. Davoli, De Queiroz, J. Cardoso, G.C. Silva, R. de Fátima, C. Moreira, A.C. Mattiello-Sverzut, Instruments to assess upper-limb function in children and adolescents with neuromuscular diseases: A systematic review. Dev. Med. Child Neurol. **63**(9), 1030–1037 (2021)
5. HHS & FDA: U. S. Department of Health and Human Services, and Food and Drug Administration, Guidance for industry use in medical product development to support labeling claims guidance for industry. Clinical/Medical Federal Register (2009), pp. 1–39

6. J. Hobart, S. Cano, Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. Health Technol. Assess. **13**(12), iii, ix–x, 1–177 (2009)
7. M.B. Jacobs, M.K. James, L.P. Lowes, L.N. Alfano, M. Eagle, R.M. Lofra, U. Moore, J. Feng, L.E. Rufibach, K. Rose, T. Duong, L. Bello, I. Pedrosa-Hernández, S. Holsten, C. Sakamoto, A. Canal, N. Sanchez-Aguilera Práxedes, S. Thiele, C. Siener, B. Vandevelde, B. DeWolf, E. Maron, M. Guglieri, J.-Y. Hogrel, A.M. Blamire, P.G. Carlier, S. Spuler, J.W. Day, K.J. Jones, D.X. Bharucha-Goebel, E. Salort-Campana, A. Pestronk, M.C. Walter, C. Paradas, T. Stojkovic, M. Mori-Yoshimura, E. Bravver, J. Díaz-Manera, E. Pegoraro, J.R. Mendell, The Jain C. O. S. Consortium, A.G. Mayhew, V. Straub, Assessing dysferlinopathy patients over three years with a new motor scale. Ann. Neurol. **89**(5), 967–978 (2021)
8. M.K. James, K. Rose, L.N. Alfano, N.F. Reash, M. Eagle, L.P. Lowes, Remote delivery of motor function assessment and training for clinical trials in neuromuscular disease: A response to the COVID-19 global pandemic. Front. Genet. **12**, 1938 (2021)
9. A. Mayhew, S. Cano, E. Scott, M. Eagle, Moving towards meaningful measurement: Rasch analysis of the North Star Ambulatory Assessment in Duchenne muscular dystrophy. Dev. Med. Child Neurol. **53**(6), 535–542 (2011)
10. A. Mayhew, M. James, H. Hilsden, H. Sutherland, M. Jacobs, S. Spuler, J. Day, K. Jones, D. Bharucha-Goebel, E. Salort-Campana, A. Pestronk, M. Walter, C. Paradas, T. Stojkovic, M. Mori-Yoshimura, E. Bravver, J. Diaz Manera, E. Pegoraro, J. Mendell, L. Rufibach, V. Straub, P.177 Measuring what matters in dysferlinopathy – Linking functional ability to patient reported outcome measures. Neuromuscul. Disord. **29**, S100 (2019)
11. A.G. Mayhew, S.J. Cano, E. Scott, M. Eagle, K. Bushby, A. Manzur, F. Muntoni, Detecting meaningful change using The North Star Ambulatory Assessment in Duchenne muscular dystrophy. Dev. Med. Child Neurol. **55**(11), 1046–1052 (2013)
12. E.S. Mazzone, A. Mayhew, J. Montes, D. Ramsey, L. Fanelli, S. Dunaway Young, R. Salazar, R. De Sanctis, A. Pasternak, A. Glanzman, G. Coratti, M. Civitello, N. Forcina, R. Gee, T. Duong, M. Pane, M. Scoto, M.C. Pera, S. Messina, G. Tennekoon, J.W. Day, B.T. Darras, D.C. De Vivo, R. Finkel, F. Muntoni, E. Mercuri, Revised upper limb module for spinal muscular atrophy: Development of a new module. Muscle Nerve **55**(6), 869–874 (2017)
13. E.S. Mazzone, S. Messina, G. Vasco, M. Main, M. Eagle, A. D'Amico, L. Doglio, L. Politano, F. Cavallaro, S. Frosini, L. Bello, F. Magri, A. Corlatti, E. Zucchini, B. Brancalion, F. Rossi, M. Ferretti, M.G. Motta, M.R. Cecio, A. Berardinelli, P. Alfieri, T. Mongini, A. Pini, G. Astrea, R. Battini, G. Comi, E. Pegoraro, L. Morandi, M. Pane, C. Angelini, C. Bruno, M. Villanova, G. Vita, M.A. Donati, E. Bertini, E. Mercuri, Reliability of the North Star Ambulatory Assessment in a multicentric setting. Neuromuscul. Disord. **19**(7), 458–461 (2009)
14. U. Moore, M. Jacobs, M.K. James, A.G. Mayhew et al., Assessment of disease progression in dysferlinopathy: A 1-year cohort study. Neurology **92**(5), e461–e474 (2019)
15. T. Morel, S.J. Cano, Measuring what matters to rare disease patients – Reflections on the work by the IRDiRC taskforce on patient-centered outcome measures. Orphanet J. Rare Dis. **12**(1), 171 (2017)
16. F. Muntoni, J. Domingos, A.Y. Manzur, A. Mayhew, M. Guglieri, G. Sajeev, J. Signorovitch, S.J. Ward, Categorising trajectories and individual item changes of the North Star Ambulatory Assessment in patients with Duchenne muscular dystrophy. PLoS One **14**, e0221097 (2019)
17. V. Ricotti, D.A. Ridout, M. Pane, M. Main, A. Mayhew, E. Mercuri, A.Y. Manzur, S. Francesco Muntoni, R. Robb, A. Quinlivan, J. Sarkozy, K. Butler, V. Bushby, M. Straub, M. Guglieri, H. Eagle, H. Roper, A. McMurchie, K. Childs, L. Pysden, S. Pallant, G. Spinty, A. Peachey, E. Shillington, H. Wraige, J. Jungbluth, R. Sheehan, I. Spahr, E. Hughes, C. Bateman, T. Cammiss, L. Willis, N. Groves, P. Emery, M. Baxter, E. Senior, L. Scott, B. Hartley, A. Parsons, L. Majumdar, B. Jenkins, K. Toms, A. Naismith, I. Keddie, M. Di Horrocks, G. Marco, A. Chow, C. De Miah, N. Goede, M. Thomas, J. Geary, C. Palmer, K.G. White, I. Wilson, The NorthStar Ambulatory Assessment in Duchenne muscular dystrophy: considerations for the design of clinical trials. J. Neurol. Neurosurg. Psychiatry **87**(2), 149–155 (2016)

18. E. Scott, M. Eagle, A. Mayhew, J. Freeman, M. Main, J. Sheehan, A. Manzur, F. Muntoni, Development of a functional assessment scale for ambulatory boys with Duchenne muscular dystrophy. Physiother. Res. Int. J. Res. Clin. Phys. Ther. **17**(2), 101–109 (2012)
19. S.C. Smith, A.A.J. Hendriks, S.J. Cano, N. Black, Proxy reporting of health-related quality of life for people with dementia: A psychometric solution. Health Qual. Life Outcomes **18**(1), 148 (2020)
20. K.I. Taylor, H. Staunton, F. Lipsmeier, D. Nobbs, M. Lindemann, Outcome measures based on digital health technology sensor data: Data- and patient-centric approaches. Npj Digit. Med. **3**, 97 (2020)
21. M.C. Walter, S. Wenninger, S. Thiele, J. Stauber, M. Hiebeler, E. Greckl, K. Stahl, A. Pechmann, H. Lochmüller, J. Kirschner, B. Schoser, Safety and treatment effects of nusinersen in longstanding adult 5q-SMA Type 3 – A prospective observational study. J. Neuromuscul. Dis. **6**(4), 453–465 (2019)

Chapter 3
Measuring Health-Related Quality of Life in Dementia

Sarah C. Smith

Abstract Dementia presents unique challenges for the measurement of health-related quality of life (HRQL). The subjective nature of the HRQL construct, the cognitive demands of questionnaires, and the necessity of sometimes relying on a proxy report taken together mean that traditional questionnaires scored using classical psychometrics may not provide robust measurement. Advanced psychometric methods, such as those based on Rasch measurement theory, can provide possible solutions to these challenges. Rasch based methods have been used with the DEMQOL/DEMQOL-Proxy disease specific HRQL instrument to measure HRQL in people with dementia, to provide robust scales, to equate self- and proxy-reported instruments, and to create a cross walk from a proxy-reported measurement to the equivalent self-reported measurement. These methods also provide a qualitative understanding of statistical change. Advanced psychometric methods such as those based on Rasch measurement theory therefore provide a potentially powerful way to address the challenges of measuring HRQL for people with dementia. However, the use of these methods is not a quick fix. They require careful development of a conceptual framework describing the construct (HRQL) and a commitment to keeping the person with dementia's perspective central at all stages of the process.

Keywords Dementia · Health-related quality of life · Equating methods · Family perspective

S. C. Smith (✉)
London School of Hygiene & Tropical Medicine, London, UK
e-mail: sarah.smith@lshtm.ac.uk

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3_3

3.1 The Challenge of Measuring HRQL in Dementia

3.1.1 The Nature of Dementia

The term dementia refers to a collection of diseases (including Alzheimer's disease, dementia with Lewy bodies, vascular and frontotemporal dementia) causing short term memory loss, difficulties with thinking (concentration, planning and organizing), language (following a conversation), orientation (losing track of what day it is, who people are) and visuospatial awareness (judging distance or seeing in three dimensions). People with dementia may also experience changes in mood, becoming more frustrated, irritable or withdrawn, anxious or easily upset or unusually sad. Dementia is progressive and degenerative and although some pharmacological treatments can delay the progress, there is currently no cure.

Globally, dementia affected an estimated 50 million people in 2017 [2] with prevalence expected to increase to 75 million by 2030. In the UK, estimated prevalence in 2020 was 907,900, rising to 1,590,100 by 2040 [62]. Dementia had a worldwide cost in 2014 of £26.3 billion [32] and in the UK alone, costs of social care for people with dementia are expected to triple by 2040 [62]. Its impact is wide ranging with effects on almost all aspects of life for both the patient [9] and their family [17].

The ability to measure health-related quality of life in dementia (HRQL) as an outcome of treatment and interventions, or to monitor the effect of disease progression, is therefore important. There have been several attempts to make recommendations about outcome measures including HRQL for use in dementia research, [59], clinical trials [57] and in routine monitoring [14, 15] with the intention that standardization of choice of instrument will improve interpretation and help to create a more meaningful body of evidence. Yet to be clinically meaningful for use in clinical trials, applied research, clinical audit and clinical practice, instruments that purport to measure HRQL of people with dementia need to be fit for purpose and provide robust and rigorous measurements [22]. To date, few disease-specific HRQL instruments for dementia have achieved this.

3.1.2 Requirements for Rigorous Measurement of HRQL

The robust measurement of a construct such as HRQL in any condition requires a number of requirements to be met. First there must be a clear definition and conceptual framework to fully describe the construct, thus laying the foundations for validity [18]. Every component of the construct should be represented by questions (or items) on the questionnaire, phrased in a standard way that is clear, non-ambiguous and easy to follow. The conceptual framework should ideally include a hierarchical continuum of the components (items) of the construct along the scale (or "ruler") [22]. When this *a priori* hierarchy of items exists, it is possible to provide *descriptions* of HRQL for any score along the scale, and thus measurement provides not only a number along the scale but also a clear understanding of what that number means [52]. This also enables change in HRQL to be evaluated in a way that is more meaningful.

The questionnaire should be completed by the person best able to report accurately on the experience that is being asked about (usually the patient themselves). Prioritizing a *self*-report in this way helps to minimize bias and further error. The questions (or items) that make up the questionnaire should be able to be combined into a scale that has robust psychometric measurement properties (reliability and validity). Typically this has been through the use of methods based on Classical Test Theory [29] and standard guidelines have been established [18, 24, 37]. These requirements create a paradox as good measurement requires that the perspective of the person with dementia themselves is kept central as far as possible, yet the nature of the condition can challenge each of these criteria. The remainder of this chapter addresses whether and to what extent each of these requirements has been addressed in attempts to measure HRQL in dementia.

3.1.3 Four Challenges to Robust Measurement of HRQL in Dementia

3.1.3.1 HRQL Is Subjective

Although there is no universal definition of HRQL, there is general agreement that it is the *subjective* impact of a person’s health condition on their life. That is, it relies on the individual’s perception and understanding of their experience of their health condition, rather than an objective report of an observable event [12]. There is obvious tension therefore with the abilities of people with dementia to articulate and express these types of concepts, particularly as insight is known to deteriorate with increasing severity of dementia. Nevertheless, international guidelines on quality of life in dementia [59] have retained this subjectivity in their definition of HRQL in dementia, remarking on “. . . . the integration of self-perceptions, a satisfactory cognitive functioning, personal activities, psychological well-being, and social interactions.”

HRQL is therefore distinct in nature from constructs of “function” (how well can a person do something) or “symptoms” (what observable indications does a person experience). For the same severity of a condition two people may report very different levels of HRQL, because they appraise their symptoms, function and health experience in a different way. This may be the result of factors such as coping mechanisms [3], available support [19], or personality [7].

Historically, frameworks that describe HRQL have typically included elements of physical, mental and social well-being [31, 60], so in addition to the subjective experience of physical aspects of health, HRQL also includes some complex components concerned with the individual’s understanding and experience of non-observable elements of health such as mental and social wellbeing. Qualitative work to explore the meaning of HRQL in dementia can be compromised because people with dementia may have difficulties with speech production and/or comprehension. In addition, the detrimental impact of dementia on reflective thinking and ultimately the loss of insight can make it difficult to ask people with dementia directly about their individual subjective experience of HRQL. Understanding what HRQL means for people with dementia and developing clear conceptual frameworks requires attentive and collaborative working with people with dementia and those who know them well to ensure that the perspective of people with dementia remains central.

3.1.3.2 Questionnaires Are Cognitively Demanding

Typical methods to minimize bias in questionnaires may not be helpful for people with dementia. Difficulties with memory, concentration and confabulation as a result of dementia can create limitations in reporting information in a consistent and meaningful manner and can mean that reporting with reference to a short time frame (such as the last week) can be unreliable. The familiar question and answer format can create challenges for people with dementia as their ability to combine cognitive functions (e.g., retention, comprehension, articulation and communication) may be reduced. Even the common strategy of using both positively and negatively worded questions to minimize reporting bias is also potentially confusing to people with dementia [46]. Likewise, although pictorial responses are preferred for people with some cognitive impairments (e.g., aphasia), they can be more demanding than a simple verbal scale for people with dementia [46]. Robust use of questionnaires with people with dementia therefore necessitates careful solutions that are grounded in consideration of the experience and ability of people with dementia.

3.1.3.3 Proxy-Report Is Sometimes Necessary

The cognitive difficulties associated with dementia mean that it can be difficult to obtain a reliable self-report of HRQL from the person with dementia themselves. Although careful questionnaire development and attention to the abilities of people with dementia mean that people with mild/moderate dementia can often successfully complete questionnaires [11, 23, 44, 45, 55], this is not possible in more severe dementia. Measurement of HRQL in dementia across the range of severity has therefore relied on a proxy (usually a family carer) to report on behalf of the patient. It is well known that agreement between patients and proxies is not always high [23, 28, 47, 51, 54], particularly for subjective, non-observable constructs such as HRQL and proxies tend to report HRQL as worse than the person with dementia themselves reports [47].

Further, in dementia there is also qualitative evidence to suggest that proxies report differences in *type* of components affecting HRQL as well as in *extent* of HRQL [44, 45]. For example, while both carers and people with dementia report an impact on the social aspects of HRQL, people with dementia describe this as a positive experience involving a social network of friends and family and the valued role they see for themselves within this community. In contrast carers tend to emphasize the negative impact of dementia on social relationships, describing unwanted or predatory social contacts and the challenges of communication in social situations. People with dementia also often compare themselves with their peers, whereas carers compare the person with dementia to how they used to be [44, 45].

3.1.3.4 Self- and Proxy-Reports Are Scaled on Different Metrics

It is not clear how to compare or interpret the different HRQL scores reported by people with dementia and those who care for them. When reported on instruments developed by Classical psychometric methods, these scores are on different scales ("rulers") and similar numbers do not necessarily have similar meanings. With relatively mild cognitive impairment people with dementia are (with appropriately designed instruments and supports) likely to be able to respond for themselves and there is little need for a proxy. As cognitive impairment progresses, people with dementia are still likely to be able to self-report but it might be necessary to also have a proxy report as carers are likely to notice different aspects of HRQL. The two perspectives are therefore complementary, though not substitutable (because they are on different scales).

As cognitive function declines further, the person with dementia is no longer able to self-report, but a family carer would be able to make a proxy report. The challenge here is in how to keep the person with dementia's perspective central even though their self-report is no longer reliable. With yet further cognitive decline, it is likely that the person with dementia is no longer living at home, and a family carer proxy-report is not appropriate because they do not see them frequently enough. Hence, the reliance on behaviorally observed instruments in later stages of dementia (Fig. 3.1). A potential solution to the proxy problem is described later.

3.2 Responses to the Challenge

Several reviews [10, 36, 44, 45, 57] have identified responses to these challenges in the form of disease-specific questionnaire-based instruments developed to measure subjective HRQL for people with dementia. Other approaches have relied solely on behavioral observation (e.g., QUALID, [58]), but as these do not assess the subjective element of HRQL we do not consider these further here. Instruments designed to measure HRQL via a written questionnaire include: Progressive Deterioration Scale (PDS) [13]; DQOL [11]; Quality of Life-AD (QOL-AD) [23]; Alzheimer's Disease

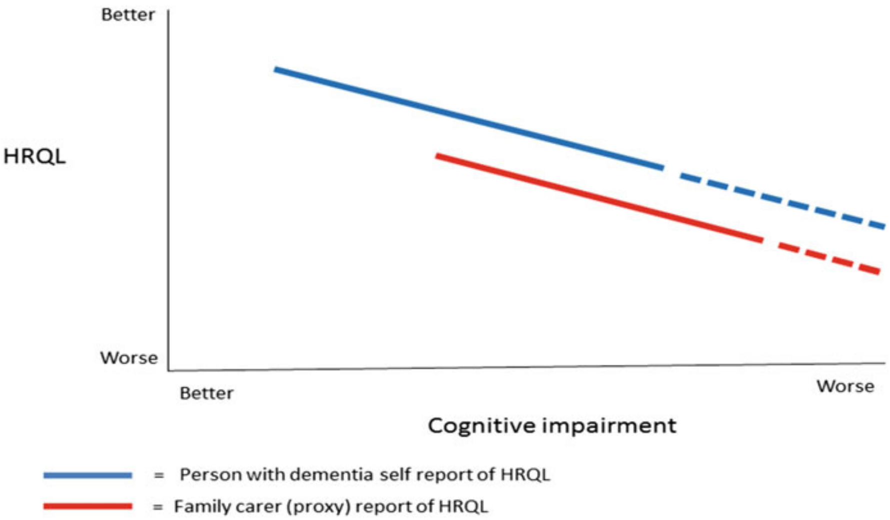


Fig. 3.1 Schematic illustration of the self- versus proxy-reporting challenge in dementia

Related Quality of Life (ADRQL) [8, 33]; Community Dementia Quality of Life Profile (CDQLP) [39–41]; The Pleasant Events Schedule – AD [1]; Quality of Life in dementia Scale (QOL-D) [53]; Cornell Brown Scale for QOL in Dementia [35]; BASQID [55]; DEMQOL/DEMQOL-Proxy [47]. A further questionnaire-based instrument has also been developed that uses individualized domains, where content is specific to each respondent (Quality of Life Assessment Schedule: QOLAS) [42], but as this cannot be used to compare individuals, is not considered further here.

3.2.1 Proxy Reported Instruments

Of these instruments, six (PDS, ADRQL, CDQLP, Pleasant Events Schedule, QOL-D, Cornell Brown Scale for QoL in Dementia) have relied solely on a proxy-report (often from a family carer). While this by-passes the reporting difficulties associated with cognitive decline in dementia and makes obtaining reliable responses potentially easier, it does not address the challenge of the known differences between self- and proxy-reports of HRQL (as described above). Even when asked to report from the perspective that they think the person with dementia themselves would give, carers often find it hard to separate their own feelings from those that they think the person with dementia has [44, 45]. These reporting problems suggest that even if a proxy–reported instrument demonstrates robust psychometric properties, it is unlikely to be an accurate reflection of the experience of the person with dementia themselves.

3.2.2 Self-Reported Instruments

Two instruments (DQOL and BASQID) have developed successful methods to elicit self –reports on questionnaires from people with dementia. Both instruments made adaptations to the method of administration to minimize the bias associated with responses from people with a cognitive impairment. Both DQOL and BASQID are interviewer administered (i.e., questions are read out verbatim by an interviewer), though responses are self-reported (i.e., the interviewer records verbatim the response given by the respondent), using large fonts and cards with the response scales printed on them, which respondents can use to point to their answer. Both instruments are reported to be reliable and valid with people with mild/moderate dementia (MMSE> = 12) [11, 55] but are not appropriate for people with severe dementia. It is therefore difficult to use instruments that rely solely on a self-report for assessments of the change in HRQL over time, as the progressive and deteriorating nature of dementia is likely to mean that there will come a point where self-report is no longer possible.

3.2.3 Instruments with Both Self- and Proxy-Reported Forms

Two instruments (QOL-AD and DEMQOL/DEMQOL-Proxy) have been developed with both self-report and proxy-reported forms. In QOLAD the same questions (13 items) are asked of both person with dementia (interviewer administered, but self-reported) and their carer (self-administered). DEMQOL (28 items) and DEMQOL-proxy (31 items) include slightly different questions but have a common core of 15 items. Originally both were developed to be interviewer-administered [44, 45], though DEMQOL-Proxy has since been found to also be appropriate for use in self-administered format [20]. It is recommended that DEMQOL and DEMQOL-Proxy should always be administered together as they address different aspects of HRQL and are therefore complementary, but not substitutable.

A key advantage of having both self- and proxy-reported versions is that in circumstances where it is necessary to take repeated assessments of HRQL (either because the outcomes of an intervention are evaluated at different time points or because the impact of disease progression on HRQL is being monitored over time) there is an appropriate reporting method for all stages of severity (i.e., self-reported for mild/moderate dementia and proxy-reported for severe dementia). However, the limitations of the psychometric methods used to develop both QOLAD and DEMQOL/DEMQOL-Proxy mean that for both instruments the self- and proxy-reports exist on different rulers (that is, the scores are sample dependent) and there is no method for determining whether people with dementia and their careers are reporting on the same construct of HRQL or whether their understanding is slightly different nor for how to combine or equate scores from the two different rulers.

3.2.4 Psychometric Approaches

All of the available HRQL instruments for people with dementia were originally developed using psychometric methods based on Classical Test Theory (CTT) [30] and although widely used at the time, instruments developed using CTT have a number of weaknesses (see [22] for an overview). Firstly, they generate scales that at best are ordinal rather than equidistant interval scales. This means the scales are inappropriate for use in many statistical analyses (because they assume interval scales) and that evaluating change over time may not be very accurate, due to the different interpretation that can be given to the semantic labels of response scales at different time points.

Thus, the response option label "sometimes" may be interpreted and used in one way by a sample at baseline but given a slightly different meaning by the sample at follow up. As there is no way to know whether this is the case for a given evaluation of change it is a hidden problem within the data which is rarely investigated within the CTT paradigm. The scores generated with CTT methods can only be used for group comparisons and not for comparisons of individual patients. This is because the measures of statistical uncertainty (e.g. the standard error) are only computed at the group level. These scores are therefore of limited use in applied clinical settings for monitoring or evaluating individual patients and even in research contexts where group comparisons might be used, there is much less confidence around the scores at the extremes of the distribution compared with those in the middle, yet there is no way of addressing this. How well an instrument performs psychometrically is dependent on the particular sample it is tested in, making it difficult to know how robust the instrument is in other samples. This makes it difficult to compare studies and challenges the understanding of how scores change over time, since these will also be from different samples. Advanced psychometric methods, such as those based on Rasch measurement theory (RMT), provide a way of overcoming these challenges.

3.3 Benefits of Using Methods Based on Rasch Measurement Theory (RMT) for HRQL in Dementia

The Rasch paradigm [5, 34] is advantageous over other approaches to measurement because the model is chosen on *a priori* grounds, rather than on the basis of whether or not the data fit the model. The model meets the criterion of invariance (i.e., that measurement should be independent of the person constructing the test and that a particular measurement should be independent of the particular items and of other people taking the test). Data that fit the model therefore also meet the criterion of invariance. If the data do not fit the model in initial efforts at calibrating a new instrument, the Rasch paradigm advocates investigating the anomalies in the data to determine why the misfit has occurred and to identify improvements that can be made in the instrument by revising items or the sampling protocol. Given a calibrated instrument with a conceptually validated construct interpretation that is shown to be stable across samples, inconsistent data patterns are no longer a threat to validity, but are instead actionable information on special strengths and weaknesses that clinicians, families, advocates, and patients can use.

In addition, in the Rasch context, combinations of items and people can be placed on the same continuum (or "ruler"), depending on the particular model that is developed. Assuming a well-developed conceptual framework and data that fit the model, this characteristic provides powerful solutions to the challenges dementia poses for measurement.

3.3.1 Diagnostic Information About the Instrument

In practical terms for HRQL in dementia, the Rasch paradigm provides a helpful set of diagnostic tools by which anomalies in the instrument can be identified (e.g. items that are not working in the way that was intended). Rather than removing these items (as might be advocated in CTT or item response theory) the Rasch paradigm provides opportunity to further investigate these items qualitatively and to determine how they can be improved. In this way, the conceptual framework of the construct being measured (i.e., HRQL in dementia) is retained and items are optimized in an iterative process to represent each aspect of the construct. This ultimately aids understanding and interpretation of scores and application to individual patients. The interval scale produced as a result of measurement developed using RMT provides greater accuracy in scores for individuals at the extremes of the distribution and provides an individual standard error, meaning that instruments that fit the model are potentially robust enough to be used at the individual level, for example in clinical decision making.

Ability to identify anomalies and to investigate why these items have not been understood in the way that was intended is particularly valuable to resolve the reporting and cognitive difficulties associated with dementia. Given a robust conceptual framework, developed in partnership with people with dementia and their carers, it is possible to retain each of the components originally deemed to be important. Items are not removed because they have not worked well in the questionnaire format (as would typically happen in CTT item reduction), rather items that misfit can be investigated to improve wording or to further hone the underlying concept. The perspective of the person with dementia themselves is therefore retained in the questionnaire and the items are expressed in a way that can be best understood by them.

3.3.2 Equating HRQL Scores

The Rasch approach provides the opportunity for a unique solution to the self- versus proxy-reporting problem. Placing both self- and proxy-responses on the same ruler (i.e., in the same model) means that it is possible to equate proxy-reports with the equivalent self-report for the same question. In this way, when it is no longer possible to obtain a self-report from a person with dementia, we can use the proxy-reports to estimate (from the Rasch model) within a stated uncertainty range what the person with dementia would have said if they were able to respond. The person with dementia's perspective is therefore not only central at the conceptual framework stage but also throughout the measurement process.

3.3.3 Quantifying and Understanding Impact on HRQL

Locating both respondents and items on the same interval continuum (or "ruler") aids interpretation of scores in a way that is not possible in traditional approaches. In addition to a quantitative estimate of change in relation to an intervention or disease progression, the Rasch approach enables a qualitative description of what each particular point on the scale means. It is therefore possible to provide qualitative description (based on the content of the items) of what a particular change means in terms of the impact on a patient's life. This is important and valuable for the practical application of such instruments in clinical decisions.

3.4 The Example of DEMQOL and DEMQOL-Proxy

Few instruments developed to measure HRQL for people with dementia have used the Rasch approach. This may reflect the spate of activity in this field in the early 2000s at a time when CTT methods were prevalent and advanced psychometric methods were only just becoming known in health-related research. Subsequently, a few studies have applied invariance scaling models to HRQL instruments in dementia in adaptations to other languages [56], evaluating instruments for new settings such as residential care [4], behaviorally based instruments [16, 38] and as part of a raft of item reduction methods to develop preference measures [25, 26]. However, DEMQOL/DEMQOL-Proxy is the only HRQL instrument in dementia to have systematically utilized the strengths of the Rasch paradigm to address the unique methodological challenges presented by measurement of HRQL in people with dementia.

3.4.1 Robust Scales for Use at the Individual Level

Analyses based on RMT [21] to establish whether and to what extent the data from DEMQOL and DEMQOL-Proxy fit the Rasch model found that Rasch measurements (based on 23 of the DEMQOL items and 26 of the DEMQOL-Proxy items) could be determined and, like all Rasch measurements, have interval properties and individual standard errors. Although future work is necessary to address the items that did not fit the model well, this approach enables greater confidence in the precision of an individual measurement and ensures they are robust enough for use with individual patients.

This is illustrated in Fig. 3.2. On the top half, the figure shows the raw scores (from the original CTT based scoring algorithm, [47]) for 3 cases referred to Memory Assessment Services for dementia (low, medium and high HRQL) and the associated error around them. Note that the error is very wide around each score and overlaps substantially between each of the three cases. Therefore, it is difficult to say whether these patients are substantially different in their HRQL. On the bottom half of the figure, the Rasch measurements for the same three cases are presented and it is clear that the error around each individual case is now much smaller, suggesting that the differences are substantive rather than part of the noise of the data.

This provides sufficient confidence in the data to potentially make it useful for clinical decision making with individual patients. For example, based on these Rasch measurements it would be possible to give patients illustration of the types of HRQL scenario that other people at the same stage have experienced and how the trajectory has developed.

3.4.2 A Solution to the Proxy Problem

RMT has also provided a practical solution to the methodological issue of proxy-reporting in dementia [49]. The developers of DEMQOL-DEMQOL-Proxy used a Rasch equating analysis to determine whether DEMQOL and DEMQOL-Proxy could be placed on the same metric (or “ruler”) and if they could, to establish a cross walk from DEMQOL Proxy to an estimate of the equivalent DEMQOL score. From the 28 items in DEMQOL and 31 in DEMQOL-Proxy, there were a pool of 44 items (of which 15 were common to both instruments). Previous analysis [21] had established that the 7 positive emotion items were not part of the same continuum and were removed from the pool, leaving 37 items (12 of which were common to both instruments).

Equating analysis was conducted on these 37 items, anchored by the DEMQOL (self-reported) items when the items were common. This model was evaluated for well-established criteria [6, 43, 61] including the extent to which there was item fit to the Rasch model; ordering of thresholds; differences in scores for different groups

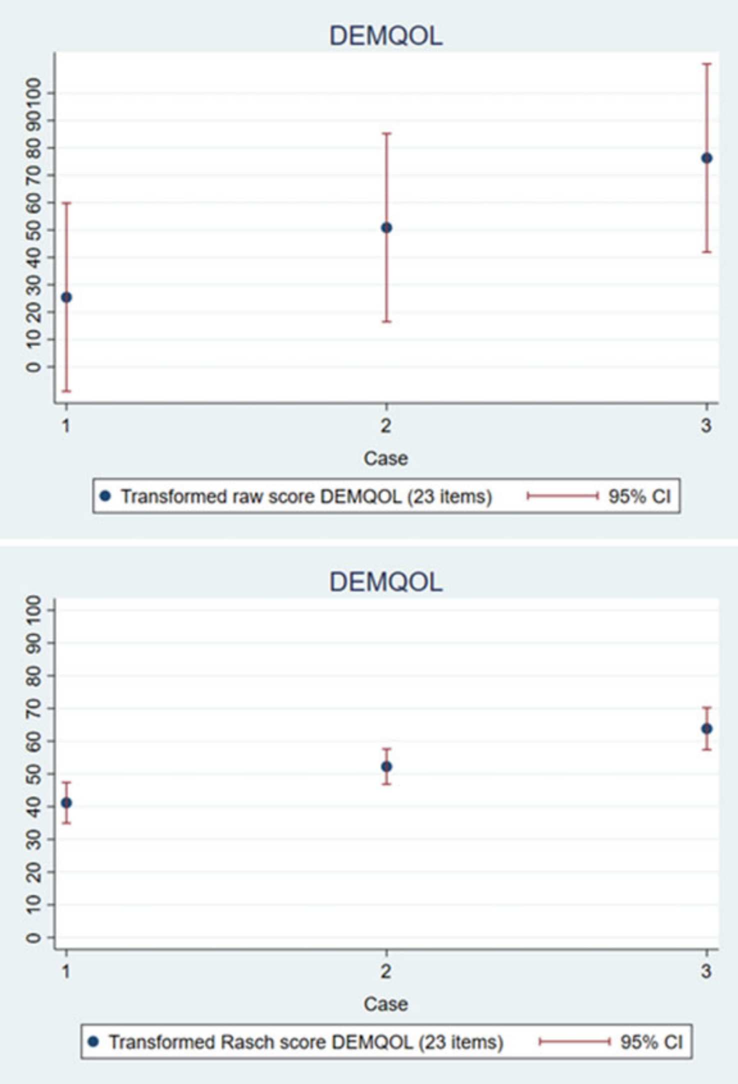


Fig. 3.2 Comparison of SE around raw scores (CTT) versus Rasch scores

(assuming the same amount of the construct being measured) (DIF); dependence of items on each other; unidimensionality and whether the items were measuring a similar range of the construct as existed in the people being measured (targeting). Results indicated that items from both DEMQOL and DEMQOL-proxy could be placed on the same metric (or “ruler”) and therefore people with dementia (reporting on DEMQOL) and family carers (reporting on DEMQOL-Proxy) were sharing a common understanding of the construct of HRQL.

As RMT places both items and people on the same scale it was then possible to estimate for any DEMQOL-Proxy score the equivalent values for DEMQOL. Cross-walk tables (see [49] mean that for every DEMQOL-Proxy score it is now possible to look up the equivalent DEMQOL score. Thus, even when a person with dementia can no longer self-report we can obtain an estimate of what their score would have been, by cross referencing the equivalent score on the cross-walk table. This avoids simply relying on a proxy-report with known imprecision and biases. This is an important development in the measurement of subjective constructs such as HRQL in dementia. For the first time it is possible to keep the person with dementia's view central throughout the process of measurement and at all stages of the disease progression.

3.4.3 Clear Qualitative Understanding of Statistical Change

The use of RMT methods has for the first time enabled a clear interpretation of the scores provided by instruments such as DEMQOL/DEMQOL-Proxy. To illustrate, imagine the case of Mr. Jones who goes to the Memory Assessment Service [48]. When he gets there, the nurse measures his height (which was 1.83 m), his weight (which was 82 kg) and he is told that there is a waiting time of 15 min. He is also asked to fill in a questionnaire and after a while the nurse tells him that his quality-of-life measurement is 67. To interpret and understand all of this information, Mr. Jones is implicitly using the standard definitions of length, mass and time and he can use that information to compare with other examples he knows about in his life experience.

For example, he understands that his height is appropriate for an adult man and that he is taller than his wife, but not as tall as the bus on which he travelled to the clinic. He understands that his weight is about average for someone of his height and that he is heavier than his young granddaughter but not as heavy as the elephant he saw in the zoo last year. He also understands that the waiting time means he will have to wait longer than it takes to make a cup of tea, but not as long as an episode of his favorite TV show. However, as there is no standard definition of quality of life, he is not sure what a measurement of 67 means for HRQL. The properties of the Rasch model, in placing both people and items on the same continuum, mean that we can help Mr. Jones to understand his measurement of 67 [48, 50].

As the measurements for DEMQOL range from 0 to 100 (where higher scores represent better HRQL), 67 is actually fairly high. The items located at that point on the DEMQOL scale indicate that this person is likely to report "a little" of a number of negative emotions and "a little" worry about not being able to do things she/he used to be able to and "a little" worry about short term memory. Someone else with a slightly lower measurement (say 56), would additionally be likely to have "a little" worry about a range of cognitive difficulties as well as beginning to feel "a little" distressed and being "quite a bit" worried about short term memory.

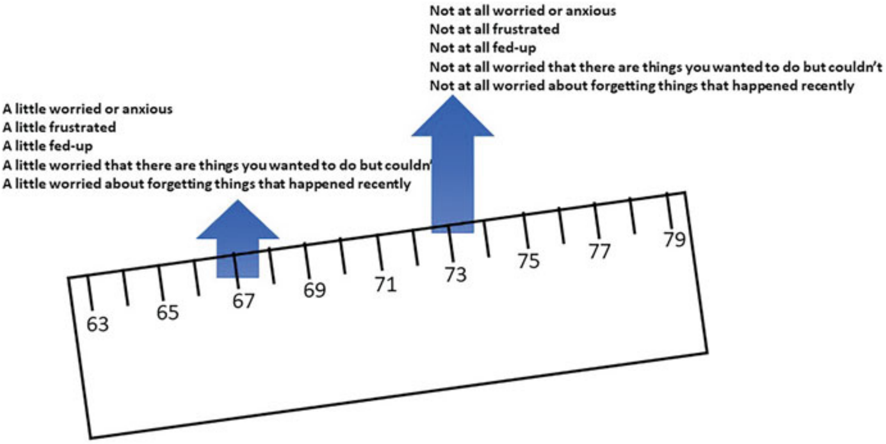

Not at all worried or anxious
Not at all frustrated
Not at all fed-up
Not at all worried that there are things you wanted to do but couldn't
Not at all worried about forgetting things that happened recently
A little worried or anxious
A little frustrated
A little fed-up
A little worried that there are things you wanted to do but couldn
A little worried about forgetting things that happened recently
63
65
67
69
71
73
75
77
79

Fig. 3.3 Illustration of how a 6 point change in HRQL on DEMQOL can be understood

In addition to these negative emotions and worry about cognitive function, a third patient who reports a fairly low measurement (say 38) would also report "a lot" of worry about the social impact of having dementia (e.g. "a lot" of worry about how they got on with people close to them, people not listening, not being able to make themselves understood). Thus we can see that on the DEMQOL scale the HRQL impact of dementia is likely to be first noticed in terms of negative emotion, as HRQL worsens impact is seen in terms of worry about cognitive function and later when impact is greatest, also worry about social impact of dementia.

Smith et al. [50] report that statistically (using distribution-based methods, [27]) the minimal important difference (MID) for DEMQOL is about 6 scale points. For Mr. Jones, a 6 point improvement (from 67 to 73) indicates that he would now be "not at all" worried or anxious, "not at all" frustrated, "not at all" fed up, "not at all" worried about things he/she wanted to do but couldn't and "not at all" worried about forgetting things that happened recently. Thus a MID change of 6 points is about 1 response option (see Fig. 3.3).

Other authors [30] have reported that people with dementia who start on anti-dementia drugs in the UK, report on average a 6-point improvement in HRQL and a similar impact was reported by people receiving psychosocial interventions (6.6 points). By providing a guide as to the most likely areas of impact, this "map" of where items are located based on RMT methodology provides a language with which patients and clinicians can potentially discuss the impact of intervention and/or disease progression on a person's quality of life.

3.5 Conclusions

The progressive and deteriorating nature of dementia presents fundamental challenges for the robust measurement of subjective constructs such as HRQL. The cognitive demand of completing self-reported questionnaires, the necessity of relying on a proxy-reported questionnaire and reliance on psychometric methods based on Classical Test Theory have resulted in instruments that have only limited precision and are not appropriate for use with individual patients. The use of psychometric methods based on RMT with instruments such as DEMQOL/DEMQOL-Proxy has largely resolved these problems. However, the use of these methods is not a quick fix. They require careful development of a conceptual framework describing the construct (HRQL) and a commitment to keeping the person with dementia's perspective central at all stages of the process. It involves working with people with dementia as partners, listening to what they tell us about what works and what is important in questionnaires. Combined with the statistical techniques embodied in RMT this provides a powerful way of improving measurement of HRQL for people with dementia.

Acknowledgments I am grateful to Dr. Jolijn Hendriks for her methodical and thorough work on the Rasch scoring of the DEMQOL/DEMQOL-Proxy instruments and Dr. Stefan Cano for his patient advice and encouragement.

References

1. S.M. Albert, C. del Castillo-Castanada, M. Sano, D.M. Jabobs, K. Marder, K. Bell, et al., Quality of life in patients with Alzheimer's disease as reported by patient proxies. J. Am. Geriatr. Soc. **44**, 1342–1347 (1996)
2. Alzheimer's Disease International webpage (2020), https://www.alz.co.uk/research/statistics. Accessed 24 Aug 2020
3. R.J. Adams, D. Wilson, B.J. Smith, R.E. Ruffin, Impact of coping and socioeconomic factors on quality of life in adults with asthma. Respirology **9**(1), 87–95 (2004)
4. E. Adler, B. Resnick, Reliability and validity of the Dementia Quality of Life measure in nursing home residents. West. J. Nurs. Res. **32**(5), 686–704 (2010)
5. D. Andrich, A rating formulation for ordered response categories. Psychometrika **43**, 561–573 (1978)
6. D. Andrich, *Rasch Models for Measurement* (Sage, Newbury Park, 1988)
7. M. Axelsson, E. Brink, J. Lötvall, A personality and gender perspective on adherence and health-related quality of life in people with asthma and/or allergic rhinitis. J. Am. Assoc. Nurse Pract. **26**(1), 32–39 (2014)
8. B.S. Black, P.V. Rabins, J.D. Kasper, Alzheimer Disease Related Quality of Life (ADRQL) User's Manual (Baltimore, DEMeasure, 1999)
9. S. Biggs, A. Carr, H.I. Irja, Dementia as a source of social disadvantage and exclusion. Australas. J. Ageing **38**(Suppl 2), 26–33 (2019)
10. A. Bowling, G. Rowe, S. Adam, P. Sands, K. Samsi, M.L. Crane, J. Manthorpe, Quality of life in dementia: a systematically conducted narrative review of dementia-specific measurement scales. Aging Ment. Health **19**(1), 13–31 (2015)

11. M. Brod, A.L. Stewart, L. Sands, P. Walton, Conceptualization and measurement of quality of life in dementia: The Dementia Quality of Life Instrument (DQoL). The Gerontologist **39**(1), 25–35 (1999)
12. M. Bullinger, R. Anderson, D. Cella, N. Aaronson, Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. Qual. Life Res. **2**, 451–459 (1993)
13. R. DeJong, O.W. Osterlund, G.W. Roy, Measurement of quality of life changes in patients with Alzheimer's disease. Clin. Ther. **11**(4), 545–555 (1989)
14. Department of Health, *The Adult Social Care Outcomes Framework 2015/16* (Department of Health, London, 2014), p. 37
15. Department of Health, *Prime Minister's Challenge on Dementia 2020. Implementation Plan* (Department of Health, London, 2016), p. 13
16. T.P. Ettema, R.M. Dröes, J. de Lange, G.J. Mellenbergh, M.W. Ribbe, QUALIDEM: development and evaluation of a dementia specific quality of life instrument. Scalability, reliability and internal structure. Int. J. Geriatr. Psychiatry **22**(6), 549–556 (2007)
17. A. Feast, M. Orrell, G. Charlesworth, N. Melunsky, F. Poland, E. Moniz-Cook, Behavioural and psychological symptoms in dementia and the challenges for family carers: Systematic review. Br. J. Psychiatry **208**(5), 429–434 (2016)
18. Food US, Administration D, Guidance for industry on patient-reported outcome measures: use in medicinal product development to support labeling claims. Fed. Regist. **74**, 1–43.16 (2009)
19. R. Gallagher, A. Sullivan, R. Burke, S. Hales, P. Sharpe, G. Tofler, Quality of life, social support and cognitive impairment in heart failure patients without diagnosed dementia. Int. J. Nurs. Pract. **22**(2), 179–188 (2016)
20. A.A.J. Hendriks, S.C. Smith, T. Chrysanthaki, S.J. Cano, N. Black, DEMQOL and DEMQOL-Proxy: a Rasch analysis. Health Qual. Life Outcomes **15**(1), 164 (2017a)
21. A.A.J. Hendriks, S.C. Smith, T. Chrysanthaki, N. Black, Reliability and validity of a self-administration version of DEMQOL-Proxy. Int. J. Geriatr. Psychiatry **32**(7), 734–741 (2017b)
22. J. Hobart, S. Cano, Improving the evaluation of therapeutic interventions in multiple sclerosis: The role of new psychometric methods. Health Technol. Assess. **13**(12), iii–168 (2009)
23. R.G. Logsdon, L.E. Gibbons, S.M. McCurry, L. Teri, Quality of life in Alzheimer's disease: patient and caregiver reports. J. Ment. Health Aging **5**(1), 21–32.4 (1999)
24. Medical Outcomes Trust, Assessing health status and quality of life instruments: Attributes and review criteria. Qual. Life Res. **11**, 193–205 (2002)
25. B. Mulhern, D. Rowen, J. Brazier, S. Smith, R. Romeo, R. Tait, C. Watchurst, K.-C. Chua, T. Loftus Young, D. Lamping, M. Knapp, R. Howard, S. Banerjee, Development of DEMQOL-U and DEMQOL-PROXY-U: generation of preference-based indices from DEMQOL and DEMQOL-PROXY for use in economic evaluation. Health Technol. Assess. **17**(5), 1–140 (2013)
26. K.H. Nguyen, B. Mulhern, S. Kularatna, B.J. Joshua, M.M. Wendy, C.T. Tracy, Developing a dementia-specific health state classification system for a new preference-based instrument AD-5D. Health Qual. Life Outcomes **15**(1), 21 (2017)
27. G.R. Norman, J.A. Sloan, K.W. Wyrwich, Interpretation of changes in health-related quality of life. The remarkable universality of half a standard deviation. Med. Care **41**(5), 582–592 (2003)
28. J.L. Novella, F. Boyer, C. Jochum, N. Jovenin, I. Morrone, D. Jolly, S. Bakchin, F. Blanchar, Health status in patients with Alzheimer's disease: an investigation of inter-rater agreement. Qual. Life Res. **15**(5), 811–819 (2006)
29. J.C. Nunnally, I.H. Bernstein, *Psychometric Theory* (McGraw Hill, New York, 1994), p. 15
30. M.H. Park, S.C. Smith, C.W. Ritchie, A.A.J. Hendriks, N. Black, Memory assessment services and health-related quality of life: 1-year follow-up. Int. J. Geriatr. Psychiatry **33**(9), 1220–1228 (2018)
31. D.L. Patrick, P. Erickson, in *Quality of Life in Health Care Evaluation and Resource Allocation*, ed. Health Status and Health Policy. Concepts of Health-Related Quality of Life (Oxford University Press, Oxford, 1993, pp 76–112

32. M. Prince, M. Knapp, M. Guerchet, P. McCrone, M. Prina, A. Comas-Herrera, R. Wittenberg, B. Adelaja, B. Hu, D. King, A. Rehill, D. Salimkumar, *Dementia UK Report* (Alzheimer's Society, London, 2014)
33. P.V. Rabins, J.D. Kasper, L. Kleinman, B.S. Black, D.L. Patrick, Concepts and methods in the development of the ADRQL: an instrument for assessing health-related quality of life in persons with Alzheimer's disease. J. Ment. Health Aging **5**(1), 33–48 (1999)
34. G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests* (Danish Institute for Educational Research, Copenhagen, 1960). (Expanded edition with foreword and afterword by BD Wright. University of Chicago Press, Chicago, 1980)
35. R.E. Ready, B.R. Ott, J. Grace, I. Fernandez, The Cornell-Brown scale for quality of life in dementia. Alzheimer Dis. Assoc. Disord. **16**, 109–115 (2002)
36. R.E. Ready, B.R. Ott, Quality of Life measures for dementia. Health Qual. Life Outcomes. **1**(11) (2003)
37. B.B. Reeve, K.W. Wyrwich, A.W. Wu, G. Velikova, C.B. Terwee, C.F. Snyder, C. Schwartz, D.A. Revicki, C.M. Moinpour, L.D. McLeod, J.C. Lyons, W.R. Lenderking, P.S. Hinds, R.D. Hays, J. Greenhalgh, R. Gershon, D. Feeny, P.M. Fayers, D. Cella, M. Brundage, S. Ahmed, N.K. Aaronson, Z. Butt, ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual. Life Res. **22**(8), 1889–1905 (2013)
38. B. Resnick, E. Galik, A. Kolanowski, K. Van Haitsma, M. Boltz, J. Ellis, L. Behrens, N.M. Flanagan, Reliability and validity testing of the Quality of Life in Late-Stage Dementia Scale. Am. J. Alzheimers Dis. Other Dement. **33**(5), 277–283 (2018)
39. S. Salek, N. Ramgoolam, S.A. Edwards, D.K. Luscombe, A.J. Bayer, in *Quality of Life Assessment in Alzheimer's Disease: Reliability of a Dementia-Specific Measure (CDQLP).* European Symposium on Clinical Pharmacy (26th) (Tours, Loire Valley, 1997)
40. S. Salek, E. Schwartzberg, A.J. Bayer, in *Evaluating Health-Related Quality of Life in Patients with Dementia: Development of a Proxy Self-Administered Questionnaire*. E.S.C.P. 25th European Symposium on Clinical Pharmacy (Lisbon, 1996)
41. S. Salek, M.D. Walker, A.J. Bayer, The Community Dementia Quality of Life Profile (CDQLP): A factor analysis. Qual. Life Res. **8**(7), 660 (1999)
42. C.E. Selai, M.R. Trimble, M.N. Rossor, R.J. Harvey, Assessing quality of life in dementia: preliminary psychometric testing of the Quality of Life Assessment Schedule. Neuropsychol. Rehabil. **11**, 219–243 (2000)
43. E.V. Smith Jr., R.M. Smith, *Introduction to Rasch Measurement* (JAM Press, Maple Grove, 2004), p. 23
44. S. Smith, D. Lamping, S. Banerjee, et al., Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology. Health Technol. Assess. **9**(10) (2005a)
45. S.C. Smith, J. Murray, S. Banerjee, B. Foley, J.C. Cook, D.L. Lamping, et al., What constitutes health-related quality of life in dementia? Development of a conceptual framework for people with dementia and their carers. Int. J. Geriatr. Psychiatry **20**, 889–895 (2005b)
46. S.C. Smith, D.L. Lamping, B. Foley, J. Murray, S.S. Banerjee, in *Obtaining Self-Reports about HRQL from Cognitively Impaired Respondents*. Presented at ISOQOL Conference (Amsterdam, 2001)
47. S.C. Smith, D.L. Lamping, S. Banerjee, R.H. Harwood, B. Foley, P. Smith, J.C. Cook, J. Murray, M. Prince, E. Levin, A. Mann, M. Knapp, Development of a new measure of health-related quality of life for people with dementia: DEMQOL. Psychol. Med. **37**, 737–746 (2007)
48. S.C. Smith, A.A.J. Hendriks, N. Black, in *Understanding DEMQOL Scores: Minimal Important Differences*. Presented at UK PROMs Conference (Oxford, 2017)
49. S.C. Smith, A.A.J. Hendriks, S.J. Cano, N. Black, Proxy reporting of health-related quality of life for people with dementia: a psychometric solution. Health Qual. Life Outcomes 18 (1) (2020)

50. S.C. Smith, A.A.J. Hendriks, S.J. Cano, N. Black, *Minimally Important Difference on DEMQOL and DEMQOL-Proxy*. Patient Reported Outcomes (Submitted)
51. K.C. Sneeuw, M.A.G. Sprangers, N.K. Aaronson, The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease. J. Clin. Epidemiol. **55**, 1130–1143 (2002)
52. A.J. Stenner, M. Smith, D. Burdick, Toward a theory of construct definition. J. Educ. Meas. **20**, 4 (1983)
53. S. Terada, H. Ishizu, Y. Fujusawa, D. Fujita, O. Yokota, H. Nakashima, et al., Development and evaluation of a health-related quality of life questionnaire for the elderly with dementia in Japan. Int. J. Geriatr. Psychiatry **17**, 851–858 (2002)
54. G. Torisson, M.L. Stavenow, E. Londos, Reliability, validity and clinical correlates of the Quality of Life in Alzheimer's disease (QoL-AD) scale in medical inpatients. Health Qual. Life Outcomes **14**, 90 (2016)
55. R. Trigg, S.M. Skevington, R.W. Jones, How can we best assess the quality of life of people with dementia? The Bath assessment of subjective quality of life in dementia (BASQID). The Gerontologist **47**, 789–797 (2007)
56. L.P. Wan, R.L. He, Y.M. Ai, H.M. Zhang, M. Xing, L. Yang, Y.L. Song, H.M. Yu, Item function analysis on the Quality of Life-Alzheimer's Disease(QOL-AD) Chinese version, based on the Item Response Theory (IRT). Zhonghua Liu Xing Bing Xue Za Zhi **34**(7), 728–731 (2013)
57. L. Webster, D. Groskreutz, A. Grinbergs-Saull, R. Howard, J.T. O'Brien, G. Mountain, et al., Core outcome measures for interventions to prevent or slow the progress of dementia for people living with mild to moderate dementia: Systematic review and consensus recommendations. PLoS One **12**(6), e0179521 (2017)
58. M.F. Weiner, K. Martin-Cook, D.A. Svetlik, K. Saine, B. Foster, C. Fontaine, The quality of life in late- stage dementia (QUALID) scale. J. Am. Med. Dir. Assoc. **1**(3), 114–116 (2000)
59. P.J. Whitehouse, Harmonization of Dementia Drug Guidelines (United States and Europe): a report of the International Working Group for the Harmonization for Dementia Drug Guidelines. Alzheimer Dis. Assoc. Disord. **14**(Suppl 1), S119–S122 (2000)
60. Study protocol for the World Health Organization project to develop a Quality of Life assessment instrument (WHOQOL). Qual Life Res. **2**(2), 153–9 (1993)
61. M. Wilson, *Constructing Measures* (Lawrence Erlbaum, Mahwah, 2005)
62. R. Wittenberg, B. Hu, L. Barraza-Araiza, A. Rehill, *Projections of Older People Living with Dementia and Costs of Dementia Care in the United Kingdom, 2019–2040* (Care Policy and Evaluation Centre, London School of Economics and Political Science, London, 2019)

Chapter 4
Improving Clinical Practice with Person-Centered Outcome Measurement

Diane D. Allen and Sang S. Pak

Abstract Measurement choices play a critical role in improving healthcare systems. As this book describes, improved measurement can promote excellence in person-centered outcomes through health policy, methodology, theory, and clinical practice. This chapter focuses on measurement choices within clinical practice that can guide decision-making for improved outcomes. The overall objective is for clinical practice to deliver high-quality and equitable healthcare tailored to the unique situation of each specific person. Separate sections discuss the context surrounding measurement choices, specific dilemmas or competing priorities that affect measurement choices, and recommendations for improving person-centered outcome measurement.

Keywords Care management · Measurement information · Competing priorities

4.1 Introduction

"Are we there, yet?" Does clinical practice achieve consistently brilliant results, yet? No. Occasional brilliance shines through: high-quality person-centered outcomes do occur [48], possibly due to brilliant people making extraordinary use of imperfect healthcare systems. These imperfect systems require improvement so that every person in healthcare gets brilliant results. With better systems, ordinarily competent clinicians and informed patients may consistently achieve excellent person-centered outcomes. Achieving excellence requires examining current practice, identifying possible barriers, and bridging remaining gaps. This chapter addresses these issues with a focus on the measurement choices made in clinical practice.

D. D. Allen (✉)
University of California San Francisco/San Francisco State University, San Francisco, CA, USA

University of California San Francisco, San Francisco, CA, USA
e-mail: ddallen@sfsu.edu

S. S. Pak
University of California San Francisco/San Francisco State University, San Francisco, CA, USA

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3_4

We begin by introducing person-centeredness and the context within which measurers make measurement choices. Three subsequent sections discuss dilemmas or competing priorities that clinicians and clinical researchers must navigate when choosing measures: personalization versus standardization, satisfaction versus effectiveness, and scientific rigor versus practical convenience. Throughout, examples promote understanding or provide evidence of improved clinical practice with person-centered measurement. Discussion includes evaluation of measurement choices that may limit interpretation of the evidence. Case A (below) illustrates the application of measurement choices when managing a multifaceted health condition. The case continues within each section to demonstrate the process of choosing measures while balancing competing interests. This chapter's final section summarizes and provides recommendations for improving clinical practice with person-centered outcome measurement.

4.1.1 Case A [8]

A person diagnosed with amyotrophic lateral sclerosis (ALS) 14 months ago meets with the multidisciplinary rehabilitation team again to make some clinical decisions. So far in this progressive and incurable disease, the person with ALS is undergoing functional losses in multiple systems. He has lost the ability to get up and down stairs, and needs help to walk on level ground. He gets short of breath after walking 10 feet. He can no longer type well enough to keep working at his software job. He has lost 10% of his body weight because of difficulty swallowing and impaired ability to manipulate a fork or spoon. What outcome measures would be most appropriate to assess his status and guide clinical practice?

Team members from different disciplines can track lab values, caloric intake, and vital capacity to document declining bodily functions; together, the team might use the ALS-Functional Rating Scale-Revised [32] or ALS-FRS-R to record the functional losses, using 12 items scored 0–4 that address different components of everyday function. From the team's perspective, this standardized scale reliably and validly identifies current capabilities; repeated measures can document the rate of decline over time [8]. From the person's perspective, the ALS-FRS-R does not ask about his home life, social support, environment, or emotional state; in short, this scale of functional deficits does not reflect the person's health-related quality of life, which can be quite variable in ALS [94, 142]. The team needs measures that address what matters most to this person. Discussions of declining function and end of life care can challenge both the patient and clinicians [46, 109]; they need measures that facilitate short- and long-term goal-setting and decision-making to optimize quality of life at each stage as the disease progresses.

4.1.2 Person-Centeredness

Person-centeredness has emerged as a critical principle underlying best practice in healthcare. No longer does the physician or other healthcare professional dictate patient management in authoritarian prescription; the individual's preferences and values take precedence; patients must be engaged as much as possible in the decision-making for their own care [36]. Further, person-centeredness, a more holistic version of patient- or client-centeredness, incorporates the total person and their life apart from their role as a patient or client. In person-centered practice, clinicians must ask people receiving care what they feel, believe, and prefer, and how they perform and respond to the activities they value in their own context, not just in a controlled clinical environment. "Asking" people can involve a number of different methods and types of measurement tools, including interview, self-report questionnaires or rating scales, performance-based measures focusing on valued tasks, biometric equipment, or wearables. Ideally, engaging the whole individual promotes attainment of the common objective: high quality and equitable healthcare tailored to the unique situation of each specific person.

Although person-centeredness underlies best practice, many individuals may never have experienced it. Healthy individuals may go to a healthcare provider based on policy recommendations: an employer or prospective travel may require a physical exam, tuberculosis test, or vaccination. The person only decides on the appointment time. When the person is very ill or has a medical condition that requires surgery or long-term care, emotional distress can affect cognition, cause anxiety, and inhibit the person's participation in decision-making. Despite that problem, clinicians and the healthcare system interested in person-centered care must diligently strive to elicit the patient's preferences. Sometimes that means involving a caregiver in the conversation or recording the information so the person can review the details after absorbing an initial diagnosis. The OpenNotes framework initiatives provide an example of greater transparency of communication when clinicians share their notes with patients, with evidence that patients find them valuable [58].

Despite occasional difficulties in extracting accurate information from patients, patient preferences contribute one of the three key tenets of evidence-based medicine or clinical practice along with research literature and clinician judgment [86]. Specifically, in evidence-based decision-making, a systematic review of high quality randomized controlled trials and the clinician's expert opinion do not take precedence over the preferences of the individual receiving care. A patient cognizant of the literature and clinician judgment may agree to proceed, negotiate modifications, request additional information or a second opinion, or decide against a procedure. Evidence-based practice presumes, however, that the three tenets work in conjunction with each other rather than in opposition, which means that each informs the others. The clinician should know the patient's preferences before examining applicable literature; researchers should have drawn from patients' experiences and outcomes in designing data collection and results reported in the literature; and the patient should be fully informed of the applicable literature, their own status, and the clinician's judgment before making decisions. The best outcome measures to advance person-centered care draw together these three components of evidence-based practice.

The United States (U.S.) formally legislated the importance of person-centeredness and all three tenets of evidence-based practice by creating the Patient-Centered Outcomes Research Institute (PCORI) through the Affordable Care Act of 2010. PCORI has since granted hundreds of millions of dollars to translational healthcare research that engages patients and clinicians at every stage: from conception to research design that includes outcomes meaningful to patients, oversight of data collection, and analysis of patient-centered changes. Further, PCORI mandates that patients contribute to every grant review panel evaluating the merit of proposed research; governmental funding agencies have followed suit in including patients in their grant review panels.

PCORI's mission, to improve outcomes important to patients, has contributed to the increasing emphasis in healthcare to keep individual patients front and center, first and always. 'Important to patients' means that assessment and outcome measurement must include not only objective measures of centimeters, grams, and number of repetitions, but also patients' direct responses on questionnaires and rating scales focusing on "patient-reported outcomes." Sometimes denigrated as subjective, in contrast to traditionally-preferred objective measures, questionnaires and rating scales have gained more universal popularity, credibility, and technical support since 2004, when the U.S. National Institutes of Health funded the creation of the Patient-Reported Outcomes Measurement Information Systems (PROMIS®). PROMIS supports development of patient-reported outcomes. It also provides a repository for standardized item banks and scales that can reliably collect patients' responses in any of 70 domains of physical, mental, and social well-being.

By 2009, the U.S. Food and Drug Administration (FDA) was publicizing Guidance for Industry [159] documents and presentations of Clinical Trial Endpoints [156] supporting patient-reported outcomes as "intuitively desirable" and "a reasonable goal of therapy." Although such outcome measures are not strictly required in FDA standards when reviewing new drugs or medical devices, these documents indicate an expectation: primary endpoints of phase 3 clinical trials should directly measure something that is important to the patient: survival, perceptible benefit, or decreased incidence of adverse events. With legislated elevation of importance and rigor of self-report measures, and impetus to consider the person's perspective throughout episodes of care, clinicians and healthcare researchers have gradually included both quantitative and qualitative measures when assessing whether interventions effectively change health or function. In addition, clinicians infused by person-centered priorities make a conscious effort to fully engage patients in all aspects of their care, from diagnosis to rehabilitation and wellness, and in all of the ways by which patients can participate in clinical decisions.

4.1.3 Context for Choosing Measures

The current emphasis on person-centeredness when choosing measures in healthcare has not evolved in isolation, but within a context that includes the clinicians and healthcare environment. Before focusing on the choice of appropriate measuring tools, the clinician and environmental contexts deserve attention.

Clinicians. Clinicians make diagnostic, treatment, and rehabilitation decisions about measurement based on their healthcare setting and what they choose to assess about the patient's healthcare needs. If one clinician treats one patient one time, then asking the patient directly for their preferences may yield the person-centered information on which to negotiate a medical decision. If the clinician sees the patient a second time, changes in the wording or circumstances of the question may result in a different response unrelated to any effects of the intervention. Keeping the words and circumstances (e.g., lighting, distractions, time since symptom-reducing medication) as similar as possible--standardizing the questions--helps focus attention on the actual treatment effects. Perhaps the clinician draws the question from the scientific literature to support evidence-based practice; in that case, changing the wording of the question may also alter the response expected. If the clinician hopes to leverage the experience with this one patient to improve practice with other patients, the clinician must standardize the question and question-delivery to minimize the effects of differences in the measure and maximize the meaningfulness of the response. Unfortunately, most clinicians do not have updated expertise in measurement along with content expertise in their specialty areas of practice. In addition, most practicing clinicians lack the time to evaluate the hundreds of new measurement tools published yearly. Thus, they may revert to common assumptions about measurement that can limit the selection of instruments and interpretation of data and thus restrict person-centered care. Five of these common clinical assumptions are described below.

Clinical Assumption 1: Standardized Measures Can Be Home-Grown and Altered.

In a 2009 study, only 218 of 456 physical therapists claimed they used standardized outcome measures; and the second most frequently employed "standardized" measures were identified as intra-facility "home-grown" measures [84]. Home-grown measures may facilitate communication within a clinical department that all agrees on the definitions and uses of that measure, but do not translate easily to people outside of that group. Further, psychometric testing for reliability of scoring and validity of interpretation and meaning frequently gets overlooked in home-grown measures. Any alterations made to standardized measures require that the clinician document the departure from the standardized measure to allow comparison across persons, testing occasions, and settings. For alterations to become standardized as in the ALSFRS-R [32] or Modified Fatigue Impact Scale [52], they, too, must undergo psychometric testing [99, 118] to ensure sound measurement.

Clinical Assumption 2. Items on a Measure Are All Equal.
Clinicians familiar with quantitative measures such as meters, kilograms, or heart rate may be tempted to treat numbers obtained from qualitative measures the same way: interpret ordinal scales as continuous measures, sum the items, and report mean and standard deviation as if on an interval scale. However, when taking a test of knowledge in a particular subject, most students would agree that some test items are harder than others. Likewise, some items on an attitude questionnaire are more difficult to endorse than others [166]. Counting the total number of items correct or endorsed does not reveal whether the easiest or hardest items are contained in that count. Examining a person's knowledge or attitude more granularly requires that measurers account for the differences among items in a measurement tool.

Clinical Assumption 3. Latent Variables Are Directly Observable.
People may assume that a score on an IQ test equals intelligence, but the responses only document how well the person reads, interprets, and marks an answer on the particular test. Constructs such as intelligence, attitude, pain, mood, and perception of one's own quality of life are not directly observable; they are latent constructs [7]. The full construct remains unknowable from the observer's standpoint; careful definition and calibration of representational measures are required to record status or effects of the construct in the person observed. In addition, the circumstances and context of the measurement dictate which aspects of each latent construct get assessed or described. Some of the differences in context include the patient's diagnosis or condition, generic measures versus those specific to specialty clinical settings (i.e., oncology versus orthopedics), and whether the measurer is a primary care provider or a specialist. Constructs such as fatigue, mood, and pain can change by the minute, hour, or day, so they require repeated measures or assessment of the impact of these person-centered latent constructs. The integrity of measuring latent variables depends on continued patient engagement; if patients do not engage in revealing this aspect of themselves, responses will be missing or inaccurate in representing the construct of interest to the observer.

Clinical Assumption 4. A Change in Score Means the Intervention Was Effective.
A change in score may merely reflect variability in the measure's ability to record the construct of interest, or variability in the person's performance due to factors such as time of day, mood, and concentration [62]. Measures such as effect size, minimal detectable change (MDC), minimal clinically important difference (MCID) [164, 167], and patient acceptable symptom state (PASS; defined as the symptom score beyond which patients consider themselves to be well) [47] can help determine if a particular change in score means that the intervention was effective and important in changing a person's life [19].

Clinical Assumption 5. Competing Measures Are Interchangeable.
Just as a change in wording may influence the respondent's interpretation of the meaning of a question, a different measure may reveal a different aspect of the construct to be measured [128]. For clinicians who recognize differences in

measures, lack of comparability can stymie incorporation of new evidence into their practice or documentation of changes a particular person experiences. Choosing which measure to use requires examination of the available evidence supporting the measures and matching measure characteristics to the appropriate characteristics of the person to be measured [35]. Equating [108] and meta-analysis [80] procedures can help interpret results across multiple measures and healthcare systems while noting the differences among them.

Clinician and clinical researcher knowledge levels vary regarding measurement assumptions and their resources differ for reviewing current literature. Even updated measurers likely compromise terminology and processes when collaborating with other stakeholders who have less-informed assumptions. In addition, measurers may choose legacy assessment tools based on their prior “state of the art” status or past endorsement; literature and practice in many areas have not yet caught up to person-centered best practices. The examples in the current chapter utilize a range of assessment tools and methodologies; caveats are provided when limitations in procedures limit the interpretation.

Environment in Which Measurement Choices Are Made. Whatever their measurement capabilities, all clinicians and clinical researchers function within a healthcare environment that can both promote and hinder best practices in person-centered measurement. Each measurer makes choices that fit their own environment and clinical setting.

One pervasive environmental factor stems from the need to apportion limited resources: prioritization of healthcare delivery can result in health inequities. The ideal of health equity seems unimaginable when pitted against the constraints of power, greed, and competing self-interests. Governmental leaders and people with wealth or the “right” mix of racial/ethnic/religious/national/gender characteristics receive healthcare that is not available to the rest of the world. Current tools and healthcare systems were created for the benefit of people with employment, health insurance, and social advantages, thus continuing systemic racism and injustices, and propagating inequities [106, 161]. Too many existing outcome measures were developed by isolated groups with a university education and extramural funding and tested on convenience samples rather than samples representing the applicable population [43]. Well-meaning measurers may be blind to a tool’s built-in cultural biases and the differences in perception that individual respondents bring to the assessment. The result is that many person-centered outcome measures may not apply to a diverse set of people or may mislead measurers in their interpretation of data [161].

For example, physical function items in questionnaires tend to assume that individuals have similar sets of daily tasks performed in similar environmental conditions. On the SF-36 [112, 163], a well-studied and common measure of health-related quality of life, two questions ask respondents whether their health limits them in walking several blocks or lifting/carrying groceries. Blocks can differ in size between urban and rural areas; the terrain can differ from clear, level sidewalks to trash-filled, pot-holed pavement; groceries may consist of a drink and chips for one or a week's supply for a family of 5. Thus, the same response to these SF-36 questions could mean widely different functional abilities. Statistical techniques such as examination of differential item functioning (DIF) can determine if groups of people, otherwise at the same level of attitude on a questionnaire, respond differently to specific items [16], but personalized differences remain unaccounted for. Cognitive debriefing after a person completes the measure might help clinicians understand patients' mindset and interpret the responses to inform clinical decision-making. Although interviewing a respondent after the questionnaire may personalize the responses, the interview also removes the advantages of brevity, consistency, and comparability that a standardized measure provides.

Clinician and environmental contexts influence measurement choices in multiple ways; conflicting expectations can add to the stress of clinical practice. The clinician may want to improve practice, but policy, access, and resource limitations may constrain measurement choices. For example, larger health organizations may have resources that enable access to references and electronic record keeping, thus lowering data collection and analysis hurdles compared to an organization with a manual entry for paper files. On the other hand, larger organizations also tend to have layers of procedural complexities in a workflow that restrict measurement choices. Further tensions arise when hospital beds or appointment slots are limited: how will measurers assess who will benefit most from care or how much care each person receives? Such dilemmas abound in clinical practice.

Questions regarding the value of uniformity and individuality also add to tension: utilizing a common terminology in measurement tools may force uniformity on persons of widely different experiences. Likewise, tools with global comparability can promote consistency across healthcare, but may also restrict individualized medicine. On the other hand, while treatment of one patient as an n-of-1 study may be very individualized, the need for economies of scale and training of new clinicians requires that pattern recognition and uniformity of measurement also develop. Each measurer must navigate such competing priorities when choosing the right measurement tools.

Healthcare administrators also face dilemmas regularly, specifically those who work in population health or quality teams alongside clinical champions, domain experts in patient-reported outcome measures (PROMs), and the information technology (IT) support team. Their priority may be to stretch the budget to pay for healthcare while showing payers that patients reliably get what they need. Such administrators may benefit from the governance of PROMs in organizations to collect PROMs more uniformly and use them to individualize care (e.g., https://epros.becertain.org/governance/guidelines/governance-structures). Other team members also must confront the challenge of minimizing resource expenditure while optimizing clinical benefit and sometimes profitability. Not all of them consider person-centered care above the financial rewards of a successful business model.

The next sections describe some of the more common dilemmas or competing priorities measurers encounter and some solutions that might help in decision-making. Examples provide common situations and evidence of clinically meaningful measurement that can improve communication and consistency in person-centered healthcare management. The ideal solutions typically balance competing priorities in a "both/and" approach.

4.2 Competing Priorities: Personalization Versus Standardization

This section discusses personalization versus standardization. Personalization tends to support person-centered care; standardization tends to support uniformity. Each has merit, but they can conflict when measurers evaluate measurement methods. This section first relates these concepts to equality and equity, then describes theoretical frameworks that can maximize the benefits of standardized measurement for equitable person-centered care.

4.2.1 Equality and Equity

Personalized medicine tailors healthcare based on the individual's pathology, genes, preferences, comorbidities, social determinants of health, and numerous other characteristics known to affect responses. Standardization of healthcare ideally means that everyone has equal access to quality care, using the same measures and interventions. However, the literature on health disparities distinguishes between equality and equity (Fig. 4.1). Equality results from all persons with the same condition receiving the same (standardized) procedures and treatment: the same surgery or the same number of rehabilitative visits. Equity results from all persons with the same condition receiving what they need (personalized) to achieve the same results: the ability to perform the same functions. Equality tends to focus on the delivery or supply of healthcare; equity tends to focus on what the person can do once receiving healthcare. This distinction has profound implications on the measures to choose for personalized and standardized care.

For example, with standardization, one unit of surgery for replacing a hip equals any other unit of the same surgery; and one physical therapy visit to improve function post-stroke is equal to any other unit of the same type of visit. This concept underlies the philosophy of healthcare reimbursement known as "fee for service," in which clinicians and institutions request payment for the number of units of different procedures delivered. Frequently identified as work relative value units (wRVUs), institutions base these on the provider's work, the institution's expenses, and the cost of malpractice insurance premiums associated with the healthcare procedures delivered. Institutions may also weight wRVUs by the severity level of the patient's condition and use them to record a provider's productivity.

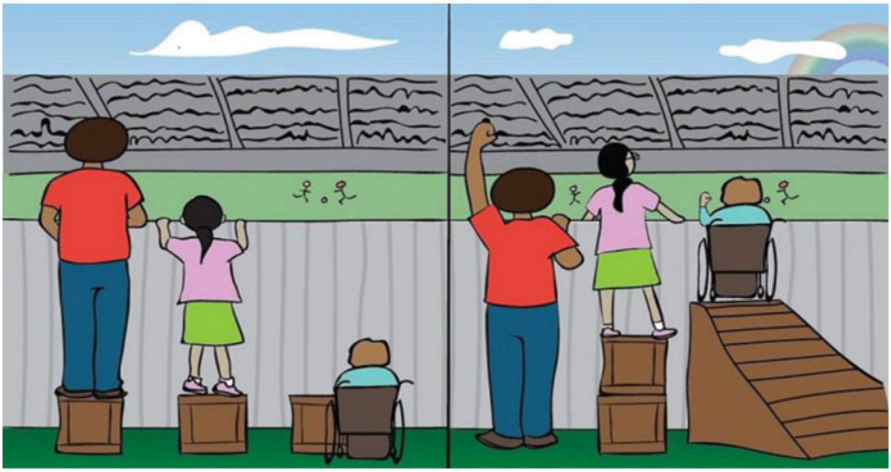


Fig. 4.1 Equality and Equity. The left image shows equal supports for all three people, but the function is not the same. The right image shows different supports that allow all three people to watch the game. Equity might also be systemically incorporated by making the fence transparent at all levels. (Image: [2])

However, while counted as the same, each wRVU disguises differences in the characteristics of all individuals involved, both providers and patients, who affect the experience and effectiveness of each unit for persons receiving care. Comparison of units (or standardized payment for units) across settings and regions will always have limitations to some degree because of differences in provider expertise and patient responsiveness.

In contrast, with personalization, differences among clinicians and patients can be addressed, with measures recording the individual patient's condition, context, and risk factors that may route the patient to clinicians with appropriate specialist expertise. Comparison of care across such personalized episodes may not reveal equality because different individuals receive different numbers of visits or procedures. However, measures of overall outcomes may reveal equity if all individuals attain similar levels of functional ability. This concept underlies the reimbursement philosophy of "value-based care," in which clinicians and institutions request payment based on persons with specific conditions achieving particular outcomes.

Note that identifying the effects of a condition for a person or population requires measurement; further, comparing overall outcomes across groups requires measurement that is standardized. Thus, equitable healthcare does not preclude standardization but mandates that person-centered outcomes form a major part of the constructs measured. Personalization then relies on grouping individuals into meaningful categories of those needing less and more attention or additional units of standardized procedures. The grouping depends on appropriate assessment of individual physical and psychological characteristics that affect outcomes from specific health conditions. The meaningfulness of the groupings determines the effectiveness of the additional procedures in contributing to equitable outcomes.

Clinical prediction rules (CPRs) can help categorize persons based on physical or psychological characteristics associated with better outcomes from prescribed intervention [111]. One example of a clinical prediction rule utilizes the 9-item questionnaire called Subgroups for Targeted Treatment (STarT) Back Screening Tool (https://www.physio-pedia.com/STarT_Back_Screening_Tool), in which patients in rehabilitation for low back pain answer Agree or Disagree that back pain has bothered them or limited certain activities, or they have worrying thoughts about it. The STarT Back Tool categorizes respondents into those with low, medium, or high risk of developing persistent, disabling low back pain. Without this tool, wide variations in patient experiences and rehabilitation have challenged resource utilization and resulted in disparate outcomes.

A meta-analysis examined studies comparing risk-stratified care based on the STarT Back tool versus standard (non-stratified) care [168]. In risk-stratified care, all patients received support and enablement of self-management. Persons classified as low risk received one treatment session; persons classified as medium risk received individualized physiotherapy with a focus on functional improvements and reduced disability; persons classified as high risk received the same plus additional training. The additional training addressed psychosocial barriers to recovery with a focus on cognitive, emotional, and behavioral responses to pain and dysfunction. In 2788 persons across 4 separate studies, applying intervention based on the STarT Back categorization compared to standard care resulted in significant and clinically important reductions in pain-related disability for each group of stratified patients; comparison of patients receiving risk-stratified care versus standard care revealed a significant although small reduction in pain-related disability plus a cost saving of more than £34 per patient.

The cost savings may reflect avoidance of unnecessary care in persons with a low risk of developing persistent, disabling low back pain [168]. Despite cost savings and positive outcomes from using the STarT Back tool, implementation of this standardized way to categorize patients with low back pain has encountered barriers [115]. In a process evaluation across 33 clinics, the rate of utilization of risk stratification was only 37.8% (range: 14.7–64.7%); author-identified barriers included staff members' knowledge and beliefs, patients' needs, technology issues, lack of clinician engagement, and lack of time [115]. Thus, standardization of a tested improvement to personalize care has had only partial success: strategies for implementing this and other improvements require additional work before being definitively recognized as improving equitable clinical practice.

Grouping patients for more equitable and person-centered healthcare must include differences in persons' previous experiences and social risk factors. *Standardized* measures might collect *personalized* information to reveal these inequities. For example, a meta-analysis reveals that homeless adults can experience accelerated aging, showing an incidence of activity limitations and fall rates usually associated with adults 4–20 years older [155]. Standardized measures across the primary studies include self-report data regarding basic (bathing, dressing, toileting, transferring, eating) and instrumental activities of daily living (taking transportation, managing medications, managing money, applying for benefits). In these studies, data were analyzed based on the number of people indicating difficulty performing at least one of the tasks listed; differences in difficulty among tasks (items) were not addressed. Fall rates were similarly compared based on self-report of fall frequency in the last year.

Frailty did not reveal significant differences in incidence between homeless and housed adults, likely because of heterogeneity in measurement choices [155]. Two studies used a 5-question survey and respondents were considered frail if they affirmed at least 3 of the 5 questions. Another study used a 42-item index of symptoms, signs, and disease classification, with a stated cut-off value that indicated frailty. The difference in granularity of the two frailty tools potentially contributed to the 4 times larger rate-ratio of frailty in homeless versus housed adults when using the 42-item index compared to the 5-question survey [155].

While additional work on measuring frailty in this population is needed, this research line explores a potentially meaningful categorization of people based on their past or present housing stability. Suppose clinicians standardize assessment by age-group, only targeting patients over 60 for routinely checking activities of daily living and fall risk. If so, they may miss physical function losses in a 45-year-old homeless individual with a premature geriatric syndrome. Similarly, for clinics to target the person-centered healthcare needs of people experiencing homelessness, resources must address physical function losses as much as psychosocial disorders. Improving clinical practice would require that past or present housing stability be considered as a risk factor for decreased functional abilities and increased falls.

Additional personalized grouping variables may either increase or decrease equitable healthcare. Population-level measures have documented differences in health and experience of pathology across height, weight, age, race, ethnic group, sexual identity and orientation, education, and insurance status. For example, in the U. S., Black patients have a higher incidence of hypertension, infant mortality, and lower life expectancy and were more likely to succumb in the first year of the COVID-19 pandemic than White patients [106]. Erroneous interpretation of these differences in health as genetic differences among races have led to appalling discrimination; inclusion of race as a factor in medical decision-making has typically resulted in under-treatment and inferior healthcare for Black patients [161]. Corrected interpretation considers such differences in health as the effects of racism, frequently resulting in crowded and unsafe living conditions, employment in essential positions that cannot be performed remotely, and delays in healthcare.

Ensuring equitable healthcare does not mean that measurers should not collect data on race, ethnicity, and socio-economic status. Examining the research on population-level differences in treatment across groups can guide improvements in clinical practice. For example, Black and Latinx patients, post-stroke, have experienced significantly greater morbidity and disability compared to White patients [54]. Armed with this data point, clinicians and patients can institute mitigating protocols and confirm their effectiveness when a repeated population measure improves. As another example, working-aged people on Medicaid (public insurance) have 14–24% lower odds of being discharged after a stroke to an inpatient rehabilitation facility and are twice as likely to get discharged to a skilled nursing facility compared to those on private health insurance plans [114]. Policymakers and administrators can use such inequities as a baseline from which to mark improvement after changes are made.

Both standardization and personalization can contribute to or detract from equitable healthcare. The choice of constructs to measure and their interpretation makes the difference. Each person collaborating in their own person-centered healthcare wants to know: “What is the effect of this proposed treatment on people like me?” Standardized measures can reveal typical effects of an intervention, but if the measures are not created or interpreted in a way that accounts for personal differences, the literature will underestimate intervention effects on one group and overestimate effects for other groups. Thus, person-centered care requires that researchers and clinicians keep equity and diversity in mind as they design and review studies with patients. Further, clinicians must discuss clinical decisions with every patient, ideally based on documented response differences in people similar to them compared to randomized controls [43].

4.2.2 *Frameworks for Assessing Effects of Health Conditions*

Theoretical frameworks can guide measurers in determining when to consider personalized variables and standardized measures in healthcare. One of the most globally recognized frameworks was drafted in 2001 by the World Health Organization (WHO): the International Classification of Functioning, Disability, and Health [79]. The ICF framework shifts the focus of medical attention from disability [121, 160] to health, thereby philosophically considering individuals and wellness as relevant constructs when mapping the course of diseases. The ICF framework allows for both standardization and personalization in healthcare across conditions, settings, and professional disciplines.

In this framework (see Fig. 4.2 showing the ICF framework for Case A), health conditions--whether wellness or pathology-- affect body structures and body functions along with the activities and participation required to engage in the multitude of roles an individual plays in life (e.g., work, leisure, self-care, relationships). Activities such as walking may become limited when function diminishes in body structures such as lower extremity muscles; participation in work or leisure activities may become restricted when walking deteriorates. The effects of health conditions are personalized by environmental and personal factors that influence how the individual responds to the impairments of body structure and function within the activity limitations and participatory restrictions that individuals encounter. Environmental factors refer to the surroundings in which activities or participation occur, including effects of physical and social conditions, and attitudes of people around whom the person functions. Personal factors refer to sex, age, coping styles, prior history with disease or disability, social background, education, and overall behavior patterns that can serve as facilitators or barriers to function. As clinical management of pathology proceeds, person-centered care requires that assessment occurs at each point of the framework to facilitate collaborative decision-making.

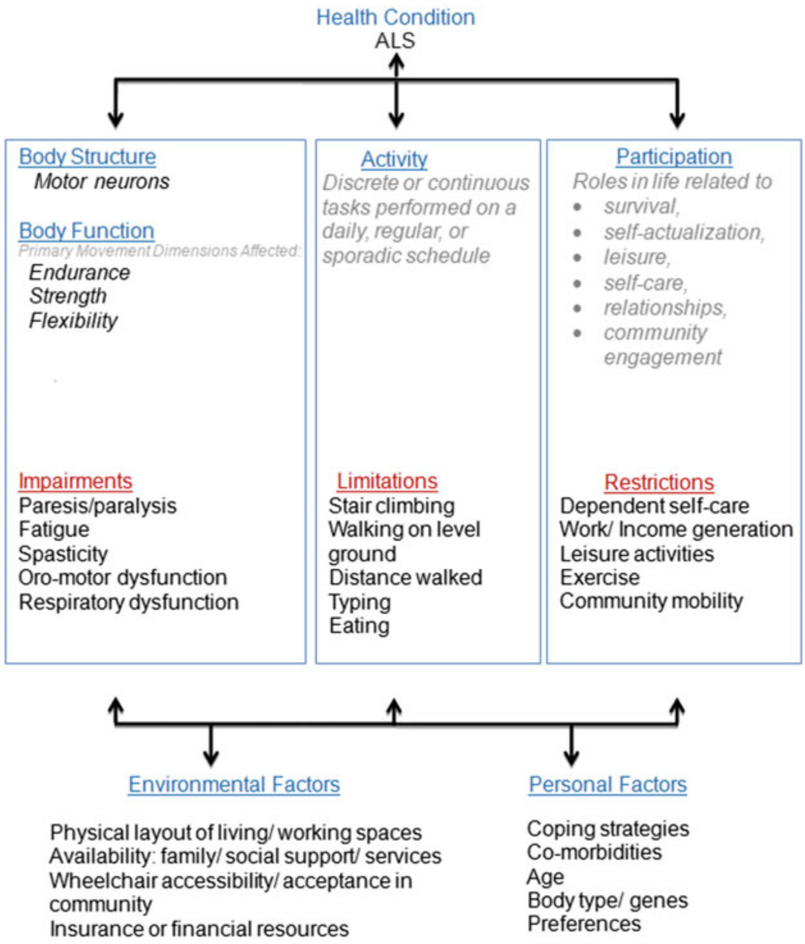


Fig. 4.2 Framework for relating the effects of ALS in Case A to the person's function based on the International Classification of Functioning, Disability, and Health. *ALS* amyotrophic lateral sclerosis. (Adapted from: International classification of functioning, disability and health: ICF. Geneva, Switzerland: World Health Organization [79])

Empirical data has confirmed the construct validity of the ICF framework; the framework components are distinct, although related [49]. Cross-sectional data were collected from 89 rehabilitation centers in 32 countries. Over 3000 persons diagnosed with one of five health conditions completed the SF-36 to enable comparison among the different conditions: low back pain, rheumatoid arthritis, osteoarthritis, obesity, or stroke. Their healthcare professionals graded the functioning of each person 0 (no problem) to 4 (complete problem) along with core sets of ICF categories (items) that had been developed previously for each individual diagnosis and across diagnoses. Multidimensional techniques were used to assess whether Rasch family models fit better with 1, 2, 3, or 4 dimensions associated with the components of body structure, body function, activities, and participation. Correlations among the dimensions ranged from 0.36 to 0.93, confirming that these components are related, but the model fit was best with all four components as distinct dimensions [49]. The findings confirmed that the ICF framework not only has construct validity but also is interpretable by healthcare professionals across different disciplines and countries.

The ICF framework facilitates clinical consideration of the interlinkages among components and the influence of contextual factors such as environmental conditions and personal factors on a person's function. These linkages can help direct a clinician in assessment and intervention planning. For example, a patient with a degenerative neurological disorder may note difficulty going to a friend's house to play bridge (restricting participation in a leisure role). The clinician can then focus assessment on the activity limitations, body structure/function impairments and personal and environmental factors that contribute to this restriction. Suppose the limiting factor is environmental, such as steps to the friend's front door or low height of the toilet seats. In such cases, interventions will involve different assessments and decision-making than if the limiting factors are inability to manipulate the cards or fatigue during an afternoon's bridge tournament. Once the direction of investigation has focused on the person's specific limitations in valued activities, standardized measures may be employed to document the person's current status and progress made with intervention.

Note that the task the person wishes to perform interacts with the environmental and personal contexts in a systems model of functional activity or movement [146]; tasks may be easier or harder based on facilitators or barriers in the environment or in personal factors. The clinician must ensure that the person is tested or asked about tasks that they want to do, at the intensity and with the stamina desired, within the sport or home/community environment where they must complete that task when participating in their preferred life roles. Personalized performance measures may either use or simulate the conditions and environmental conditions for the tasks of interest; self-report measures can ask patients what impact their health condition has on performance of valued tasks that are either pre-specified or patient determined.

The ICF framework and the systems model underscore potential differences in effects that a similar health condition might have on various people. Two people with the same amount of lower extremity weakness may continue working for different lengths of time, depending on the job requirements (sedentary or active), employer's support, aptitude and liking for the work, and the financial incentive for maintaining it. As another example, two people may report knee pain from identical pathology, but their pain experience does not simply indicate the threat or real-time occurrence of tissue damage [66]. Instead, the pain experience reflects each person's assessment of how dangerous, intense, and frequent the pain is, and the effect of pain on their activities and participation. Successful person-centered healthcare relies on consideration of the person's tasks, the environment in which they perform those tasks, and the personal factors of anxiety, depression, fear, belief system, and prior experience that affect how they view any activity that induces pain [66].

Thus, the ICF framework and systems model together provide a theoretical foundation for personalizing healthcare while utilizing standardized measures to assess targeted components. The standardized measures can contribute needed gradations for prescribing treatment and comparing outcomes across time and persons.

4.2.3 *Case A, as Informed by the ICF Framework (Fig. 4.2) and Systems Model*

In the initial description of this case, the patient with ALS is meeting with the rehabilitation team, indicating the team's interest in engaging the patient's preferences along with clinician judgment in person-centered decision-making. The ICF health condition in this case is ALS. The pathology and identified activity limitations suggest that impairments of body structures and functions include neurological structures and function, along with progressive weakness in all extremities and some oro-facial muscles. The patient can no longer participate in his prior work roles and has become more dependent in the role of self-care; other participation roles and restrictions have not been specified. The ALS-FRS-R assesses the activity component of the ICF framework, indicating the severity of limitations in activities of daily living and some restrictions in his participation in self-care.

The literature shows that quality of life for people with ALS varies, depending more on perceptions of control and communication than actual severity of the disease [94, 142]. To improve person-centered care, the team will need a greater understanding of the patient's additional participation roles and restrictions to know what tasks the patient values, and the contextual factors related to person and environment. The choice of measures should focus on the quality-of-life indicators of importance to the patient, standardized to grade goal attainment or further restriction, personalized to apply within the patient's context. Critical decisions regarding use of a feeding tube or ventilatory assistance must be addressed at some point; the team and patient will need to consider the implications for extension of life versus quality of that life. Such decisions must take into account the person's attitude and circumstances as well as progression of the disease.

4.3 Competing Priorities: Satisfaction Versus Effectiveness

This section contrasts satisfaction and effectiveness. Satisfaction means contentment with the process or a particular status. Effectiveness means the person made progress toward some goal. Person-centeredness does not in itself dictate sole assessment of either construct. However, measures of satisfaction do not substitute for measures of effectiveness when preferred outcomes include more than satisfaction with the experience. This section first differentiates between these constructs, then examines several measures related to person-centered effectiveness: biomarkers and wearables, disability, discomfort, and personal factors that include perceived quality of life.

4.3.1 Distinguishing Satisfaction from Effectiveness

Gauging satisfaction and person-centered effectiveness both likely require some version of self-report questionnaires. The distinction comes in the questions asked and the comparisons made. Instruments assessing satisfaction typically focus more on a customer/business relationship than the effects of the intervention and are completed at the end of the individual's experience. Questions about available parking, helpfulness of staff making the appointment, and cleanliness of the restrooms hopefully lead healthcare businesses to improve the next customers' experience with regard to infrastructure, personnel, or maintenance. Healthcare businesses, like other businesses, may want high net promoter scores, meaning that patients indicate more top ratings than low ratings on questions about likeliness that the person will recommend this business to others [3].

On the other hand, unlike retail or food service businesses, healthcare businesses may be financially penalized for repeat visits, particularly if the return (e.g., to the emergency room or re-hospitalization within 30 days) indicates previous ineffectiveness of care or premature discharge for the same health condition. Thus, in addition to interest in patient satisfaction with the interaction, healthcare businesses also have an interest in the effectiveness of interventions. Questions about patient-perceived effectiveness might relate to curing or minimizing an adverse condition, slowing degeneration, stabilizing or improving activity, and increasing participation in valued life roles. Such questions presume that a condition or activity level has changed; thus, the measurement must occur at least twice, for a baseline and post-intervention data point. Differences in responses across repeated measures could lead clinicians to improve clinical practice: identifying gaps in practice, implementing best practices, and engaging patients in shared decision-making. Measurers interested in effectiveness must differentiate what construct they value and wish to improve, then measure in a way that can show status and change in that construct. Most satisfaction measures include very few items related to effectiveness and their single time point does not address change. Measuring effectiveness may require separate tools that specifically address the constructs that the clinician and patient hope will change with intervention.

Cott et al. [38] attempted to merge the constructs of satisfaction and person-centered effectiveness by developing a client-centered measure of perception of care received for clients who had undergone inpatient rehabilitation services. The authors defined their concepts and the component domains through literature review, focus groups with clients, and review by content experts. The seven resultant domains included: participation in decision-making and goal setting, education, evaluation of client-centered outcomes, family involvement, emotional support, coordination/continuity, and physical comfort. The authors drafted five-six questionnaire items for each domain and tested item statements for clarity and relevance through cognitive interviews with patients; all items had response choices on a 5-point Likert scale with 5 being strongly disagree and 1 being strongly agree. The 33-item Client-Centred Rehabilitation Questionnaire (CCRQ) was then tested for internal consistency, test-retest reliability, and discriminative construct validity in a mailed survey to patients discharged from two rehabilitation facilities. Cronbach's alpha and test-retest reliability proved acceptable based on the authors' stated standards, and the subscales differentiated as expected among patients with different primary diagnoses [38].

In evaluating the resultant CCRQ, the measure's success is mixed with regard to capturing satisfaction and effectiveness. Only one of the 7 CCRQ domains asks about outcomes, with 4 statements for respondents to endorse [38]: "I was kept well-informed about my progress in areas that were important to me"; "I accomplished what I expected in my rehabilitation program"; "The program staff and I discussed my progress together and made changes as necessary"; and "I learned what I needed to know in order to manage my condition at home." Other categories of questions focused more on whether respondents felt they were treated well, not especially if they were treated effectively: shared clinical decision making, inclusion of family, and whether valued symptoms and conditions were addressed in a personalized manner. Although a literature review examining the CCRQ confirmed the 7 domains as important in person-centered rehabilitation literature [131], a German study conducting an exploratory and confirmatory factor analysis could not confirm the original 7 dimensions of the CCRQ [95]. Körner et al. [95] proposed a 3-factor structure across 20 items: decision-making/communication; self-management/empowerment; and psychosocial well-being. Their proposed 20 and 15-item versions no longer include the item "I accomplished what I expected in my rehabilitation program," the one item that comes closest to answering whether the respondent got better with treatment. Another study of a slightly modified CCRQ [51] concluded that person-centered goal setting (determining which outcomes mattered) varied based on numerous factors: respondents agreed (90%) that they were encouraged to participate in goal setting but agreed less strongly (73%) that decisions about what would help them were made in collaboration with program staff [31]. In general, the literature advocates use of the CCRQ for assessing whether a rehabilitation program is person-centered, and what aspects of the program might require improvement to increase person-centeredness, but the CCRQ is not specifically an outcome measure of intervention effectiveness.

A person's satisfaction with care could conceivably influence the effectiveness of that care. Studies have shown relationships between satisfaction and the likelihood of adherence to treatment [77]. In other words, people who express satisfaction with care may also adhere better to their clinicians' recommendations. Further, if barriers in access to the facility or the person's communication with program staff prevent a person from receiving or following through with recommendations, then reduced effectiveness may follow. Satisfaction may also relate to health-related quality of life

[77], a common construct typically associated with the ICF's participation component and measured to assess restrictions that might ensue from the particular health conditions. However, characteristics that influence a person's positive or negative outlook on life can also influence responses to questionnaires about satisfaction and quality of life. The positive or negative bias on these measures further restricts their usefulness in conveying information about the effectiveness of healthcare. In a systematic review of studies examining patient satisfaction with musculoskeletal physical therapy, a meta-analysis across 7 of the 15 studies revealed that patients generally marked "satisfied" to "very satisfied" with care, with clinician's attributes as the most consistent determinant of patient satisfaction [77]. Only 3 of the 15 studies identified treatment outcome or symptom improvement as important to satisfaction. The authors of the review concluded that positive response bias may have affected outcomes, and overall satisfaction was more related to interactions with the therapist and the process of care than the outcome of the treatment. The implication is that any measure of effectiveness likely requires multifactorial assessment to distinguish between symptom improvement and satisfaction with overall care [77].

Clinicians may hesitate to put effectiveness to the test: measuring satisfaction seems less risky because patients might be satisfied with their experience even if the healthcare has not been effective. In addition, clinicians may fear that patients know what they want but not what they need; "allowing" patients their say requires trust that they will receive, apply, and appreciate recommendations even when not exactly the solutions that patients thought they wanted. Researchers advocate transparency to build greater bilateral trust: clinicians must elicit relevant information from and listen to patients, share information about the health condition, and collect relevant data regarding the patient's status. Clinicians must then ensure that patients see the data indicating their own changes with intervention. With guidance to interpret visualized data, patients can determine for themselves when their outcomes have improved. When clinicians and patients both agree that person-centered outcomes show improvement, bilateral trust may ensue.

4.3.2 Person-Centered But Not Person-Reported: Health and Disease Biomarkers

Health and disease biomarkers can help measure healthcare effectiveness. They are person-centered when they measure a construct important to the patient and patients can see the data themselves. Usually these instruments record fluctuations in body structure or function while a person engages in meaningful activity; as such, biomarkers can function as a type of biofeedback. Biomarkers such as heart rate, blood pressure, and body weight can help assess health, wellness, and disease progression. Heart rate and blood pressure can be ascertained through biometric equipment that ranges from Intensive Care Unit monitoring to smart-watches. An overall increase in blood pressure monitors can mean that the person is working hard or straining homeostatic systems; a decrease might mean that the body is more relaxed, has greater vascular efficiency (as in trained athletes), or is about to pass out from decreased oxygen to the brain (as in orthostatic hypotension). Thus, biomarkers must be monitored and interpreted before using them for clinical decision-making; patient education regarding the fluctuations in these measures can help individuals use the data more effectively in controlling their own activities.

Walking speed has emerged as an effective biomarker for multiple conditions; some researchers call it a functional "sixth vital sign" [55]. Gait velocity, or distance covered in designated time, is measured across various distances and with equipment as simple as a stopwatch or complex as instrumented gait mats or computerized motion analysis systems. The starting instructions may specify either self-selected gait speed or "as fast as possible while staying safe." Slower gait velocities have been associated with fall risk, functional decline, hospitalization, and discharge destination in several populations. Gait velocity is a person-centered measure in the sense that slow velocities can limit activities such as walking across a busy street before the light changes. Cut-off values have been proposed to indicate that a person walking at velocities below 0.4 m/s likely has ambulatory ability restricted to household use. A community ambulator typically requires a gait velocity of 0.8 m/s or greater [55]. Interventions that can progress gait speed from one category of function to the next make an important difference in a person's life.

Physical activity or step count as recorded by wearable equipment continue to evolve as potential biomarkers of disease progression or intervention effectiveness [22]. Research grade and commercially available technology typically employ accelerometers with carefully calibrated algorithms to count steps or movements, usually summed and averaged by the day. The advantage to wearable equipment is that clinicians and patients do not have to rely on one-time snapshots of walking behavior in the clinical environment to make decisions; they can monitor and assess behavior changes during everyday participation in life roles [23]. For everyday use to become feasible, however, devices must be inexpensive enough for each individual to have one, and person-friendly enough for individuals to wear with any clothing options, and remember to resume wearing after recharging the batteries. Person-centeredness argues for commercially-available technology worn continuously long-term rather than a returnable device worn for a few days or a week [23]. Persons in many diagnostic groups have used step count monitors for healthcare and personal data collection [22]. In multiple sclerosis (MS), continuous step count monitoring adds value to clinical assessment because the wide range of step count levels within each level of disability on the clinician-reported Expanded Disability Status Scale (EDSS) [98] can potentially alert patients and clinicians to disease progression or goal attainment more quickly [24].

Nonlinear Variability When measuring biomarkers continuously, it is tempting to characterize the results with an overall mean and standard deviation. However, recurring bodily functions such as heartbeat, postural sway, and stepping during gait occur dynamically, varying over time: the next heartbeat, postural sway direction and amplitude, or step length and width does not exactly match the previous one [74, 90]. Patterns of variation have importance and can be measured [150]. For example, postural sway forward and back and side to side can be recorded as the displacement path of the person's center of pressure while standing quietly on a force plate [67, 75, 150). The pattern of repetition from one forward to backward sway cycle to the next is neither robotically repeating nor randomly varying, but normally depicts a certain amount of complexity necessary for "optimal movement variability." This Goldilocks pattern ensures that the body is not too rigid to respond quickly to changes in muscle contraction or environmental alterations and not too variable to accomplish the task of staying upright. Nonlinear measures of structural variability over time include approximate entropy (a measure of unpredictability) [126, 127] and Lyapunov exponent (a measure of divergence) [73, 75, 149], among others. Optimal movement variability is person-centered in the sense that individuals function from a particular growing and aging body structure within a changing environment that presents various obstacles to movement. From this perspective, variability can present information about the state of the body and environment from which to make adjustments, not simply errors in movement program selection or execution. Each individual must then adapt each goal-oriented movement appropriately to accomplish the task as desired. Skill does not mean freedom from errors but greater variability so the individual can flexibly adapt.

Hunt et al. [75] demonstrated a critical implication that supports the examination of optimal movement variability in person-centered care. In a health condition such as MS, problems with balance and excessive postural sway can result in falls and fear of falling. Traditional reports of postural sway have used linear measures of center of pressure displacement such as range and root mean square. However, patterns of postural sway over time can present as either abnormally rigid (low approximate entropy or Lyapunov exponent) or abnormally random (high ApEn or LyE) in their variability compared to the movement patterns of healthy controls. People with abnormally rigid patterns might respond best to interventions that introduce variability: perturbation training or supported practice correcting induced errors in balance. People with abnormally random patterns might respond best to interventions that restrict variability: repeating the same task while constraining part of the movement or augmenting sensory feedback. Mixed groups of patients receiving the same intervention might wash out differences in both the linear and nonlinear measures of variability, resulting in the erroneous conclusion that the intervention had no effect [75]. Meaningful categorization of persons is necessary to reveal the actual effectiveness of some interventions, and then to apply the appropriate intervention that matches the needs of each person. Nonlinear measures of variability can distinguish important differences that may guide clinicians in determining when to repeat the same task to minimize error and when to provide variations similar to those the person encounters daily to practice adjusting responses.

4.3.3 Person-Centered Disability as a Measure of Effectiveness

Measuring disability depends on the definition of the term and purpose of the measurement [10]. The classic medical model of disability [78] defines it as the effect of trauma, congenital factors, or disease. In such cases, the purpose of measuring is to assess how different the person is from normal health or function. A disability may be defined as the inability to work because of a medical condition that is expected to last one year or more or result in death [147], and measures would determine who will receive disability insurance benefits. Measurement based on traditional medical or governmental definitions can serve business or societal purposes but typically lack appropriate information to guide person-centered clinical practice. For the latter purpose, "meaningful disability" may be defined as that which affects abilities most critical to a particular patient's health-related quality of life [141]. Measures would then focus on patients' preferred bodily functions, activities, and participation roles, as the ICF components that patients most want to change.

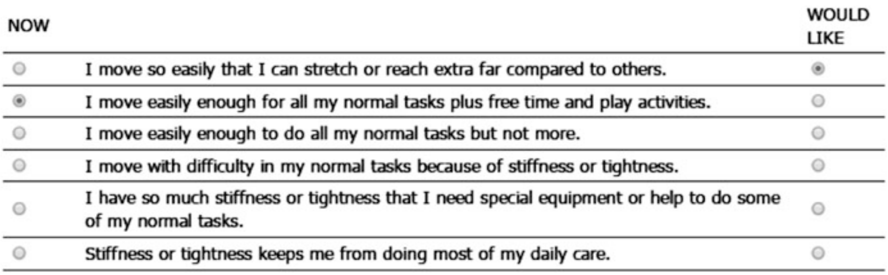
To define meaningful disability, Mitra proposed that "an individual is disabled if he or she cannot do or be the things he or she values doing or being" [119]. When aligning this statement with the ICF framework, "doing" implies bodily function applied toward the performance of an activity; "being" implies the roles the individual assumes for participation; and what the individual "values" is a personal factor that influences preferences and satisfaction when engaging in activities in the individual's environment. From the individual's perspective, disability is thus the gap between doing and being, between having and wanting with regard to the ability to do or be. Measurement of meaningful disability must include both the having and wanting components [10]. While assessment of current ability the individual has can be accomplished with performance-based or disease impact measures, assessment of preferred ability that the individual wants requires observation of behavior choices during daily living or self-report to determine what values the individual places on various functions.

Standardized functional assessment tools ask respondents to record their current ability across tasks that the creators or clinicians deem valuable in activities of daily living; the patient's preferences typically do not get concurrently assessed. For example, most functional measures include walking items, but wheelchair athletes may prefer the speed and agility of their wheelchairs to slow, clumsy and non-functional walking with braces. Common measures of disability also focus on what people can do currently, using "normal" function as the standard by which to gauge the effects of the health condition on the individual [50, 63, 83, 98]. Measures that assess self-perceived disease impact come closer to addressing the doing-valuing gap Mitra identified [119]. For instance, an MS impact scale asks respondents how much their disease has affected their ability to do specific everyday tasks [70]. With this phrasing, respondents compare current abilities to their own remembered normal rather than to a hypothetical societal norm.

Another way to examine effectiveness related to disability is through quality-adjusted- or disability-adjusted life years (QALYs or DALYs, respectively). QALYs indicate the effectiveness of medical treatment regarding increases in the length and quality of life the patient experiences. The quality-of-life estimation has sometimes come from the clinician's perspective, but association with patient-reported quality of life measures ensures the person-centeredness of the QALYs [134]. A systematic review of studies basing QALYs on patient-reported quality of life measures before and after treatment indicated that almost half of the 70 studies in their review had what the original authors deemed acceptable cost per QALY. The authors of the systematic review qualify their findings because of differences in QALYs when based on differences in the quality-of-life measures used, and the variability of costs per QALY in different regions [134]. In contrast, disability-adjusted life years or DALYs indicate the number of life years while living with the effect of a disease or disability [33, 57]; thus, instead of years and quality of life gained as recorded in QALYs, DALYs focus on the years of life lost and years spent in ill-health [14, 33]. People's perceptions of the effects of disability differ among countries and populations. In one developing country, healthcare providers and mothers deem severe and profound disabilities as far worse than death; mothers who have a child with a disability that has required hospitalization deem even moderate disability as worse than death [148]. Factors influencing these differences include societal stigmas, environmental accommodations for particular disabilities, access to rehabilitation, and financial support. Measuring with QALYs or DALYs assumes that all persons value life years that are free of disability, but that people might tolerate a certain amount of disability if healthcare could extend the total years of life. However, attaining the individual person's perspective requires individual-level measures that show what they value most in their lives.

The literature provides several options for aligning individual abilities with the person's preferences. The Patient-Specific Functional Scale (PSFS) [154], Goal Attainment Scale [92], and others [101] utilize goal setting to indicate patient preferences [76]. The patient chooses the tasks toward which the patient works in rehabilitation in consultation with the clinician. Both measures show evidence of reliability and validity [20, 76, 88, 105, 151, 152]. The patient-generated tasks in both measures facilitate repeated measures in individual patients at the beginning and end of care but limits the usefulness of each measure for comparing across patients. Further, neither scale explicitly records gaps remaining at the end of an episode of care between outcomes achieved and the patient's preferred abilities. A Movement Ability Measure was generated that addresses these deficits [6].

Movement Ability Measure. The Movement Ability Measure was developed based on measurement principles [165] to operationalize Mitra's [119] latent construct of disability specifically for rehabilitation. The Movement Ability Measure assesses self-reported current and preferred movement capabilities [7] based on the Movement Continuum Theory (MCT) of physical therapy [37]; the interval logit (log of the odds) scale allows calculation of the resultant current-preferred gap. According to the MCT, the focus of therapy is to minimize the gap between preferred and



NOW		WOULD LIKE
○	I move so easily that I can stretch or reach extra far compared to others.	●
●	I move easily enough for all my normal tasks plus free time and play activities.	○
○	I move easily enough to do all my normal tasks but not more.	○
○	I move with difficulty in my normal tasks because of stiffness or tightness.	○
○	I have so much stiffness or tightness that I need special equipment or help to do some of my normal tasks.	○
○	Stiffness or tightness keeps me from doing most of my daily care.	○

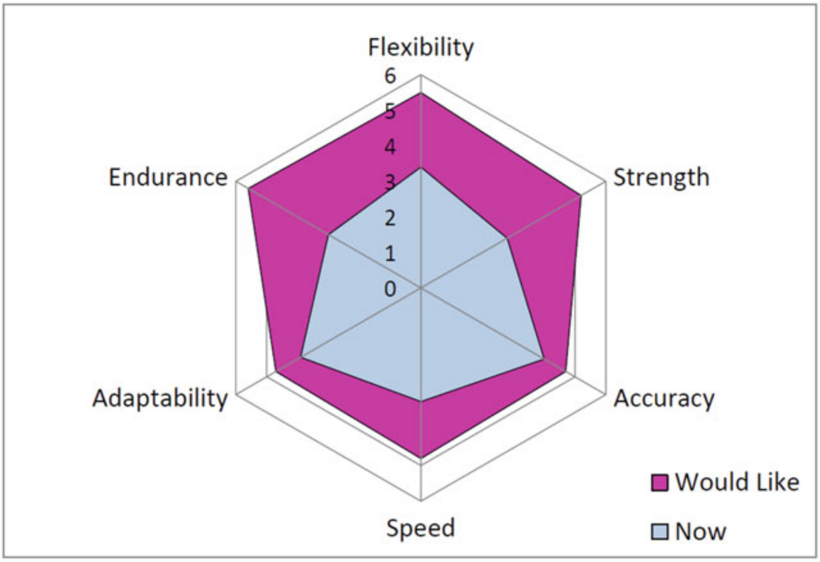
Fig. 4.3 One item on the Movement Ability Measure associated with the movement dimension called flexibility. (Adapted from Allen et al. [12])

current abilities within the person's maximum capabilities [37]. To extend the MCT and facilitate testing of the theory, Allen [4] subdivided movement into six dimensions: flexibility, strength (force production), accuracy (including timing and direction), speed, adaptability (ability to make dynamic adjustments to movement based on sensory input), and endurance [4]. The six dimensions evolved from clinical practice, literature review, and the criteria needed for robot development; each dimension subsumes related terms to result in a set of dimensions that fit predetermined criteria: descriptive, efficient, distinct, measurable, and understandable. Once the dimensions were identified, structured interviews of prior patients informed the clinical hypotheses about behaviors and performers that align with less and more ability to move. The 24-item Movement Ability Measure was generated using six ordered statements regarding self-perceived current and preferred movement abilities for each item, with four items in each dimension. The six statements or response choices for each item target abilities across the range of the construct from less to more (Fig. 4.3).

Testing reveals the strong link retained among the Movement Ability Measure, the six-dimensional movement ability construct, and the extended Movement Continuum Theory [7]. When testing "current ability" responses to the Movement Ability Measure, the one-parameter multidimensional model aligning items with their appropriate dimensions fit best, and 52% of the 318 community-dwelling respondents showed differences rather than a uniform average across dimensions [4].

The Movement Ability Measure also shows evidence of reliability, content and construct validity [6], and responsiveness of the current ability responses to physical therapy intervention [5]. Examination of the current-preferred gaps indicated that respondents discriminated between current and preferred abilities, and most indicated a preferred ability other than the top statement for each item [9]. Patients starting physical therapy had a larger gap than healthy controls, and their perceived gaps significantly narrowed after 2 weeks in physical therapy [9].

Testing also indicates that this person-centered outcome instrument can improve clinical practice. A computer-adaptive test version of the Movement Ability Measure (MAM-CAT) was developed and used for a two-part pragmatic study [144]. The first part of the study assessed differences in patient priorities as revealed by the



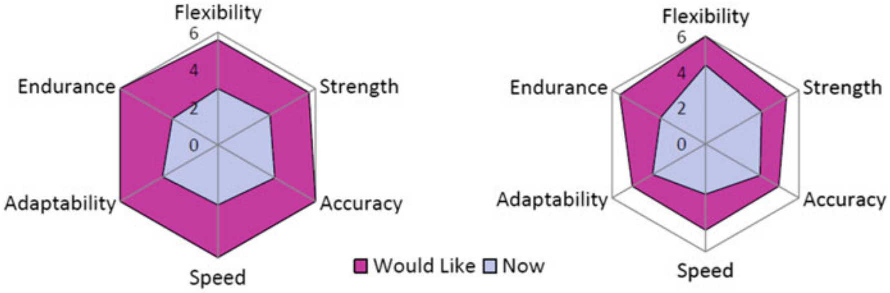
	Now	Would Like	Gap
Flexibility	3.4	5.5	-2.1
Strength	2.8	5.2	-2.4
Accuracy	4.0	4.7	-0.7
Speed	3.2	4.8	-1.6
Adaptability	3.9	4.7	-0.8
Endurance	3.0	5.6	-2.6
Total	20.3	30.5	-10.2

Scale	Movement Ability Levels:
5.1 - 6+	Moves Competitively
4.1 - 5	Completes Normal Activities Plus Extra
3.1 - 4	Completes Normal Activities
2.1 - 3	Moves with Difficulty
1.1 - 2	Needs Help to Move
0 - 1	Cannot Do Even With Assistance

Fig. 4.4 Movement Ability Plot (MAP) from the self-report MAM-CAT depicting the gaps between the movement ability the person in Case A likely has *now* and the ability the person *would like* to have. Dimensions with the largest gap (endurance at −2.6, with strength and flexibility closely following) reflect the person's priorities for care. The original logit scale has been transformed to a scale in which 0 means cannot do even with assistance, and 6 means moves competitively. (Figure adapted from Allen [8])

movement dimensions showing the largest gaps on the MAM-CAT (Fig. 4.4) and clinician emphases as revealed by the health records of assessments and interventions during the episode of physical therapy [12].

Although both the MAM-CAT and physical therapy notes indicated that patients progressed, comparison showed poor or slight agreement on the movement dimensions to prioritize (Fig. 4.5). The second part of the study compared the first results with patient outcomes when MAM-CAT responses were reviewed and discussed between therapist and patient at the beginning of the episode of physical therapy [13]. The average decrease in total gap size at the end of care was significantly greater when therapists and patients had the opportunity to view the MAM-CAT responses together.



Initial visit:
PT focus was flexibility and strength; patient indicates endurance and speed have the biggest gaps.

Upon discharge:
Patient notes improved flexibility and strength; patient indicates endurance and speed still have the biggest gaps.

Fig. 4.5 One patient with a diagnosis of low back pain in which the emphases from the physical therapy notes do not match the movement dimensions with the largest gaps according to MAM-CAT results. Flexibility and strength improve by discharge, and the current-preferred gaps narrow, but the largest gaps remain in endurance and speed at discharge. (Adapted from Allen et al. [12])

4.3.4 Person-Reported Effectiveness: Minimizing Adverse Conditions

Satisfaction and effectiveness may come closest to merging when the health condition targeted by healthcare is self-perceived pain, fatigue, dizziness, other discomfort, or fear. From the patient's perspective, the adverse condition is the reason they are seeking healthcare; effectiveness in managing the condition will likely weigh heavily in their satisfaction. These conditions, however, can change rapidly: an overall improvement with intervention may be masked by a current spike in discomfort. Thus, a person-centered assessment must consider the multiple facets of the targeted condition so that both the immediate condition and the impact of the condition on everyday life can be revealed.

Numeric or visual analogue scales have been used to assess the level of a person's current pain, fatigue, exertion, or dizziness. Numeric scales typically use the form: "on a scale of 0 to 10, with zero meaning no pain and 10 meaning the worst possible pain, what number would you say your pain is right now?" Some numeric scales have qualifying phrases attached to several of the numbers rather than just at the ends of the scale; Borg's Rating of Perceived Exertion has respondents put a number on how hard they think they were working during the target activity. On a 6–20 scale, 7 is very, very light, 13 is somewhat hard, and 19 is very, very hard. The scale has been shown to correlate with exercise heart rates, with 15 on the scale approximately equal to 150 beats per minute [26]. A category ratio scale is also used, with numbers 0–10 indicating increasing exertion. Visual analogue scales generally have patients mark the distance on a line or a more severely wincing face that represents their discomfort. Studies have shown that a VAS measure of the impact of fatigue on daily life has moderate reliability and can rapidly screen individuals for severe fatigue impact on their life [96].

The impact of pain on function has been assessed in well-documented measures such as the Roland-Morris Disability Questionnaire [139] and the Oswestry Low Back Pain Disability Questionnaire [50] for patients with low back pain. Dizziness has been assessed more generally in the Dizziness Handicap Inventory [81], which can be used across diagnostic groups.

Fatigue measures have been devised in general or for specific use in a particular diagnostic group. For example, the 9-item Fatigue Severity Scale (FSS) [97, 103, 117], has been used across multiple populations with neurologic disorders, including MS, stroke, and ALS [59]. Unfortunately, researchers disagree on the scale's usefulness; in a Rasch analysis of the FSS in people post stroke, 2 of the 9 items did not fit their model and a 7-item FSS showed better psychometric properties [103]. In contrast, a Rasch analysis of the FSS in people with MS showed that 4 of the 9 items did not fit a unidimensional construct of the social impact of disease; the authors recommend using a 5-item version [117]. Some measures subdivide fatigue into cognitive and physical dimensions [118, 157] and others report that factor analyses do not support such a distinction on the MFIS [132]. Hudgens et al. [72] developed the Fatigue Symptoms and Impacts Questionnaire—Relapsing Remitting Multiple Sclerosis using the best available guidance regarding instrument creation and testing with Rasch and exploratory factor analyses. The resulting instrument has multiple types of items to address both current severity and impact. The authors report good reliability and validity that align with the patient-derived constructs from which the measure was generated [72].

4.3.5 *Person-Reported Mediators and Measures of Effectiveness*

Personal contextual factors such as the person's self-efficacy or confidence, and their coping capacity, mood, happiness, or quality of life may reveal mediating variables that affect the person's responses to treatment [137], or may be targeted as outcomes themselves.

For example, perceived self-efficacy as recorded in scales such as the 14-item Self Efficacy Scale for exercise/physical activity provides respondents with various adverse conditions that might conceivably hinder them from partaking in the activity [56]. Examples of three items include "I could exercise … when tired; when my schedule is hectic; when I haven't reached my exercise goals." The level of a person's self-efficacy may mediate their ability to change behaviors such as smoking, over-eating, or a sedentary lifestyle [122].

As another example, stronger coping capacity as documented by the 13-item Sense of Coherence scale has been linked with greater health-related quality of life in healthy populations and people with diseases such as ALS [142]. An example of a mediating variable that may also become an outcome in itself is confidence in one's ability to perform various tasks without falling, e.g., the Activities-Specific Balance Confidence scale [130] which is both related to fall risk and an indicator of reduced fear of falling. Likewise, health-related quality of life may mediate the impact of the person's health condition on their body structure, function, and activities, or may become the outcome that healthcare specifically targets.

Health-related quality of life measures may apply specifically to those with a particular diagnosis or across diagnoses, as the commonly cited Sickness Impact Profile (SIP) [21] and SF-36 do. The SIP is a 136-item questionnaire covering 12 categories that patients might perceive to have been impacted by their adverse health condition: (a) sleep and rest; (b) emotional behavior; (c) body care and movement; (d) home management; (e) mobility; (f) social interaction; (g) ambulation; (h) alertness behavior; (i) communication; (j) work; (k) recreation and pastimes; and (l) eating. SIP scores are represented as percentages of disease effect.

In contrast, the SF-36 incorporates the concept of wellness along with disease impact into the range of quality of life. The SF-36 is a 36-item short form of a much longer health-related questionnaire in the Medical Outcomes Study; the eight sub-scales include "eight of the most frequently represented health concepts" [163, p. 906]. The sub-scales consist of general health, physical functioning, role physical, role emotional, social functioning, bodily pain, vitality, and mental health. In addition, one question asks about the person's health transition, comparing their health to one year ago. Multidimensional analysis of the functioning of the SF-36 across diagnostic groups in the U.S. [112] confirms the 8-dimensional construct and hypothesized order [133] of the items and response levels. Treatments that change physical morbidity have the most effect on sub-scales relating most to the physical summary scale, while treatments that target mental health mostly affect sub-scales correlated more highly with the mental summary scale [163].

Comparison of the SF-36 across populations and with scales specific to particular diagnoses can inform researchers of characteristics of the scales. For example, SF-36 responses in people with ALS are generally lower than in age-matched controls across all dimensions except for bodily pain which is nearly the same in the two groups [94]. However, functional deficits as recorded in the ALS-FRS-R are significantly associated with deficits in the physical functioning and vitality subscales of the SF-36, not the other aspects of health-related quality of life. In contrast, increasing depression as recorded in the Beck Depression Inventory is associated with all SF-36 subscales except for physical functioning. Although not an inevitable consequence of incurable disease, moderate to severe levels of depression have been identified in 30% of people with ALS [94].

4.3.6 Case A, as Informed by Satisfaction and Effectiveness Measures

In the initial description of this case, the patient with ALS is meeting with the rehabilitation team. The team prioritizes the person-centeredness of care and has each patient complete the Client-Centred Rehabilitation Questionnaire (CCRQ) at the end of an inpatient stay to gauge the team's success in that realm. Based on the underlying constructs of person-centeredness, the team emphasizes seven domains: patient participation in decision-making and goal setting, education, evaluation of person-centered outcomes, family involvement, emotional support, coordination/continuity, and physical comfort. In the team meeting with the patient, they discuss the person's social and home environment, lifestyle, and priorities. They determine together which family members and healthcare specialists will participate in the next team meeting, the information the patient and family still need before making upcoming decisions, and the outcomes and mediating variables most important to assess.

Recommended mediating and effectiveness measures include continuation of the ALS-FRS-R, and inclusion of the SF-36, MAM-CAT (see Fig. 4.4), Sense of Coherence (SOC) scale, and Beck Depression Inventory. Assessment of current-preferred gaps in movement ability with the MAM-CAT will help determine the patient's current priorities for addressing body function impairments. Assessment of coping capacity (with the SOC) and depression will help determine the usefulness of psychotherapy as part of the multidisciplinary emotional support. The most relevant adverse condition currently impacting the patient is fatigue; the Fatigue Severity Scale will be added to the outcome assessments to document changes with instruction on movement efficiency or use of energy-saving devices. Education regarding the use of a feeding tube or non-invasive versus tracheostomy mechanical ventilation at home will incorporate assessment of QALYs to assist the patient and family in analyzing the pros and cons of different decisions [138]. In addition to the measures recommended here, each discipline will use their own specific instruments. The team will need to consider the patient-burden of all these measures going forward.

4.4 Competing Priorities: Scientific Rigor Versus Practical Convenience

This section contrasts scientific rigor and practical convenience. In the ideal world, stakeholders in healthcare all have scientifically rigorous and technically sound measures at their fingertips to meet all their measurement needs. However, as much as clinicians and researchers desire scientific rigor, constraints in time, personal knowledge, availability, and clinical usability remain determinants when choosing instruments. Practical convenience, therefore, must be considered so that the average clinician treating both the everyday and unusual patient has access to brilliant tools to promote brilliant results. This section briefly reviews types of criteria associated with rigor in measurement, then examines ways that measurers have attempted to make measurement more convenient, accessible, and usable.

4.4.1 Ideal Measurement

Ideally, developing scientifically rigorous tools involves collaboration between those who see the need for a construct to be measured and those who create and assess the instrument. Best practice for generating new person-centered measures requires collaboration among patients and clinicians to define the construct of interest and base the questions or tasks on theory, so the items represent a scale of the construct as intended (for examples, see the CCRQ, MAM-CAT, and the Hudgens fatigue measure already described in this chapter). Metrologists and psychometricians can ensure that the resulting items fit the underlying construct and function as designed. Complex constructs may require complex instruments with multiple dimensions or question types to better align with the underlying theory [6, 72]; measurement experts can both develop and test such complexity using factor analysis and multidimensional procedures. Building instruments based on item response and Rasch measurement principles ensures that differences in an item or task difficulty factor into the assessment of people's abilities or attitudes, and each individual respondent gets located on a unidimensional or multidimensional construct with an individualized standard error of measurement. Psychometricians can ensure that resultant measures reduce the uncertainty around the individual's measurement.

Sometimes multiple published measures already assess some portion of the construct of interest, such as functional abilities while in an acute hospital or in outpatient clinics but not across settings [64]. Measuring a construct using a single metric across settings has person-centered advantages because patients (and their clinicians, if the settings share data) who have seen their data in one setting can track their progress as they move to the next setting in an episode of care. Instead of developing a measure from scratch, measurers might pool relevant items from multiple measures, hypothesize the order of the items on the construct, and test all of the items in a sample representing the population to be tested. Measurers can thereby create an item pool. For example, the PROMIS database banks items across many different constructs. Using Rasch measurement techniques, measurers can determine the usefulness and difficulties of individual items and response sets and refine the pool.

Once an item pool has been generated, the next step likely involves creating subsets or groups of items because fewer items take less time and decrease patient burden. However, indiscriminate cutting of items in a pool or on an established test can diminish reliability and sensitivity [11]. One technically complex solution uses computer adaptive tests, in which the computer program adapts the test by choosing which items get presented to the respondent from the item bank [144]. The programmed algorithm selects the next item based on responses to prior items; the objective is to fine-tune the location of the respondent on the latent construct.

Another way to expedite test-taking is to create fixed sets of items as short forms, using items that span the range of interest on the construct so that floor and ceiling effects are minimized in the population for whom the measure is created. For example, the Activity Measure for Post-Acute Care (AM-PAC), based on the activity component of the ICF framework, originated by gathering a pool of 222 items across 8 original sources/instruments. The creators wanted clinician- or patient-reported short forms for each of three domains: physical/movement, applied cognition, and daily living [64]. They tested items relevant to each domain with a Rasch one-parameter model, and whittled down to 10 items in each domain, with overlapping sets for inpatient and community (outpatient) use, all on the same metric. The originators simultaneously created a CAT version [65]; the CAT version has the advantage that estimates for individual scores are more precise than when using the fixed short forms. Subsequently, a version for acute care, the AM-PAC 6 Clicks mobility measure [85], has been widely used to alert providers to functional needs and predict discharge disposition [124].

When clinicians do not have access to item pools for the constructs they want to measure, they may want to understand how one measure compares with other measures, particularly measures that address the same or similar construct. Equating methodologies can inform stakeholders of overlap and differences among various measures, using common items to help calibrate items that are distinct on multiple measures to ascertain what areas of the latent construct are well-covered [108].

When equating methodologies have not yet compared measures of interest, measurers may look to the experts to ascertain the best measures to use in research or practice. Systematic reviews and meta-analyses gather related published evidence to compare the usefulness of one intervention or measure over another in a particular population. The advantage is that meta-analyses can gather data across measures of a similar construct; when combining controlled trials of interventions, meta-analyzers typically use Cohen’s d, which expresses the difference in means in standard deviation units [80]. The drawback is that many of the measures in these meta-analyses are reported in and analyzed as continuous scales despite their ordinal scale origins. Readers of these studies should interpret findings with caution.

Published clinical practice guidelines [29, 120], consensus standards, or recommendations [89, 110, 123, 129] can direct readers to the measures that have the most literature support to date. Their advantage is that the review typically comes from a panel of content experts, covers much more literature than the individual practitioner has available, and summarizes designated aspects of the various measures. Their drawback is that the recommendations do not commonly guide clinicians in choosing the right measure for a particular patient [35]. Also, panels typically choose the aspects of reliability and validity they will review and prioritize measures with the most literature, so newer, less-utilized measures may receive lower recommendations despite the use of more rigorous development criteria and better alignment with the latent construct.

While guidelines and recommendations generally specify the need for continual updates, resources may not keep up with the rapidly changing literature. Additional resources may come from task forces that review sets of measures and attempt to standardize the best set for research [44] or practice (e.g., a consensus statement on core outcome measures for persons with Covid-19 released by the American Physical Therapy Association, June 2020) in particular populations.

While the judgment of experts matters, the clinician still must make an individual choice for each patient. Assessing effectiveness requires that the selected outcome measures be responsive to the changes expected with the proposed intervention. If the clinician expects the intervention to improve or delay deterioration in function, then the function must be measured using tools that are sensitive to detect changes from the intervention. Further, if the patient wants to get better at function, then the measure should assess functional ability on the most interesting tasks to the patient. While many different statistics have been used to assess instrument responsiveness [5], those that align change with patient-assessed clinical importance or quality of life likely have a better person-centered focus [140]. The types of statistics can include minimal important difference, minimal clinically important difference, or PASS.

4.4.2 Accessibility and Convenience

With a scientifically rigorous and responsive measure or recommendations in hand, the measurer then requires access to the instrument and ease of use for each patient. Access to the instrument may include fees paid to use a copyrighted instrument along with convenient methods for patients or subjects to answer the questions or perform the tasks, clinicians or researchers to collect and enter data, processors to analyze data and interpret results, and clinicians to share visualized, interpretable data with the patient. Convenience is critical because different patients will need different sets of instruments, so measurers cannot make measurement choices just once, but newly for every diagnostic group and patient.

Large data sets can assist in making decisions even at the level of the individual patient because they show associations and results that can be expected with various treatments. The next two subsections discuss data repositories and electronic health records, and the promises, challenges, and future possibilities in relation to person-centered outcome measurement.

Data Repositories. The National Library of Medicine defines a data repository as “a place that holds data, makes data available to use, and organizes data in a logical manner” [42]. Just as person-centered care must attend to a patient’s multifaceted characteristics and context, a system that holds person-centered data must provide for multifaceted data types collected from various sources. Unlike historical use of a data repository as a siloed, single storage unit limited in data sharing, the concept of “Information Commons” [158] fosters collection from multiple data sources such as electronic health records (EHR) and biomedical research data to collaborate and accelerate research, advance clinical care, and improve person-centered outcomes [15, 25, 162]. For example, Donado et al. developed a pediatric pain data repository integrated with EHR data to collect longitudinal patient-reported outcome measures [45]. Using the 6 core areas recommended by the International Associations for the Study of Pain, the authors implemented validated questionnaires from EHR to capture patients' quality of life (e.g., physical function, emotional functioning, sleep, mood) and assess their progress and response to treatments [45].

The rehabilitation domain has traditionally been slow to adopt data repositories despite the potential benefits when addressing shared challenges in rehabilitation research. Challenges include rarity of some health conditions which limits sample sizes for determining evidence of efficacy and effectiveness [30], and the "black-box" phenomenon meaning that variability in symptom presentation hinders characterization of the mechanism of effectiveness of interventions or treatments used across research studies [68, 82]. Other challenges include inconsistent use of standardized outcome measures [84], and insufficient collection of pertinent patient attributes (e.g. comorbidities, gender, age, geocodes, etc.) that mediate their response to treatment.

However, early stages of the development of data repositories in some populations denote notable progress in meeting these challenges. A Multiple Sclerosis Rehabilitation Repository (MSRehabRep) can potentially store, retrieve, and share rehabilitation information in MS to support clinicians and researchers [27]. Focus groups from various stakeholders prioritized key features to "envision" retrospective and prospective data use in the MS population [27]. Desired characteristics for the repository included security and accessibility for data use and data sharing, and most essential, high-quality data standards [27]. Further, addressing the ongoing challenges of establishing common data elements and standardized outcome measures across diverse MS populations was an important step identified in data repository development [27].

The American Physical Therapy Association (APTA) has created the Physical Therapy Outcomes Registry [125] to collect clinical data from therapists on a registry platform, a type of data repository. A registry is defined as "an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves one or more predetermined scientific, clinical, or policy purposes" [60]. Registry data has several key advantages for researchers and clinicians. Along with providing opportunities for comparative effectiveness research [153], the combined use of clinical EHR data and other uniform data of interest (i.e. patient outcomes data relevant to a specific population) can establish benchmarking outcomes data and fulfill regulatory quality reporting requirements [125]. Desired characteristics overlap between data repositories and the PTOR features: enabling commonly used patient-reported outcome measures, establishing a core set of de-identified patient data, and collecting longitudinal data from an episode of care in physical therapy [125, 153]. While enrolling in PTOR is optional for users, organizations that currently participate in quality payment programs (e.g., Merit-based Incentive Payment System or MIPS) or other quality initiatives will likely benefit from using registries like PTOR because the Centers for Medicare & Medicaid Services have designated it as one of the Qualified Clinical Data Registries to collect and report on behalf of therapists [125]. This designation makes the PTOR particularly valuable if an organization does not have Office of National Coordinator (ONC)-certified EHR systems nor sufficient resources to collect, manage, and report outcomes data from their native EHR systems.

Ongoing hurdles hinder the utilization of data repositories to accurately reflect person-centered care. First, patients and their caregivers should know that having a data repository means that their data could appear online, even if de-identified, which may not preclude deduction of personal data if few people have the target condition. In addition, a clinical repository of data will have less ability to help direct person-centered care if it does not clearly identify the patient's priorities and values. Thus, patient engagement is needed throughout the development cycle of a data repository as well as future data use. Second, strong data governance is needed to address privacy and data sharing policies and procedures. Even with de-identified patient data, HIPAA privacy rules and data sharing agreements need to comply with federal and local organizational guidelines. Third, development and maintenance costs could be a challenge for smaller health systems; they may need to consider outsourcing solutions to third party entities or clinical registries. The outsourcing itself can present challenges when attempting to select from several options without deep knowledge of the criteria to optimize. Fourth, curating different data sources to ensure that a repository remains reusable and standardized while continually updating with new clinical data types could be very resource-intensive [162]. Fifth, despite current cloud computing power and its robust nature to handle big data, barriers to accessibility and data use hinder administrators, researchers, and clinicians, let alone patients, from taking full advantage of refined and standardized data to inform clinical decisions.

As healthcare reimbursement practices pivot from fee-for-service to value-based care, harnessing person-centered data to reduce variations in practice, along with any unnecessary treatment costs, becomes non-negotiable. Data repositories in rehabilitation can provide an entry point to critically reevaluate person-centered rehabilitation, potentially uncovering ways to improve clinical practice at an aggregate level. However, to discover and deliver value-based care for our patients, healthcare needs a robust system that facilitates the collection of the right patient-reported outcomes for the right types of patients so that providers and patients have the ability to access the relevant data to make shared decisions. The fundamental groundwork of standardizing core data elements and outcome measures must continue if data repositories will succeed in promoting the use of person-centered outcomes to improve clinical practice.

Electronic Health Record Systems. The Federally funded Health Information Technology for Economic and Clinical Health (HITECH) Act incentivized adoption of electronic health record (EHR) systems and the Affordable Care Act (ACA) promoted their interoperability so that patients might "carry" their health records from one provider to another. These changes represented an evolution of person-centered data collection and processing methods across US healthcare.

Even so, the integration of patient-reported outcome measures (PROMs) with EHR systems has been moderately slow, partly due to EHR vendors' priorities to meet the early stages of Meaningful Use (MU) regulatory requirements. Specifically, under Stage 1 and 2 of MU requirements, the emphasis was primarily on adopting EHR systems, capturing and advancing process measures, and promoting interoperability, not explicitly prioritizing PROMs.

However, from stage 3 of the MU requirements to the Quality Payment Programs in Merit-based incentive payment system (MIPS), the focus shifted to evaluate "person and caregiver-centered experience and outcomes" [69]. This shift mandated the collection of PROMs to achieve MIPS quality measure requirements. Concurrently, different specialty practices were shifting toward implementing PROMs in real-time clinical care, including oncology, orthopedics, and pediatrics as well as primary care practices [17, 18, 45, 145]. Increased usage of PROMs put additional pressure on vendors to include common PROMs in their EHR systems. Currently, a growing number of health organizations have implemented two of the largest EHR vendors in the market, Epic Systems and Cerner Corporation, and benefited from its certified EHR systems integrated with a suite of PROMs such as the PROMIS measures.

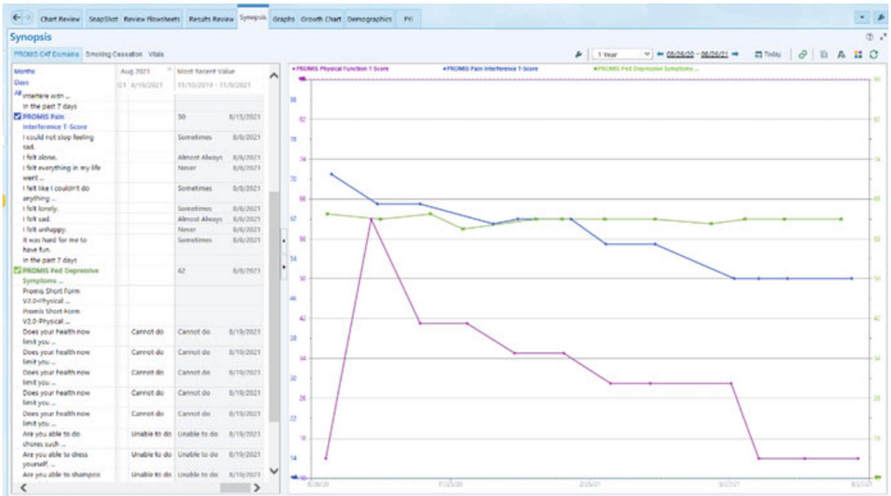
Several essential areas continue to evolve with the integration of PROMs in EHR technology: accessibility for rapid point-of-care data collection and visualization as well as remote administration of PROMs, and data integration into routine clinical management. The objectives of data integration into routine management is for the clinician to be alerted to changes and patient priorities, and clinicians and patients to review person-centered data together so that the measures play a part in shared decision-making [12].

Timeliness and convenience of electronic patient-reported outcomes (ePRO), when present, facilitate data completeness and utilization for aligning practice with person-centered changes. For example, using a patient portal tethered to the EHR system allows clinicians or staff to assign (or automatically assign based on preset organizational rules such as diagnosis, visit types, language, etc.) and send ePRO questionnaires securely to patients. Patients then have the flexibility to complete their assigned questionnaires at their convenience through available communication devices such as a web-based portal, smartphone app, or kiosk.

Specifically, PROMIS CAT questionnaires can be made readily available and exchanged electronically through a patient web portal. As soon as the patient completes a questionnaire online, it gets processed in real-time and allows assessment of results to be routed to the care team or specific referring clinician. If further action is warranted due to severe symptoms, then automatic triggering of notifications to different care teams or clinicians allow them to make necessary referrals and medical decisions in a timely manner [28].

For instance, if a new patient in an outpatient physical therapy clinic hits a cutoff point for depression in the PHQ-9 (9- item Patient Health Questionnaire), then the electronic notification may help alert the clinician of the need for psychiatric consultation. However, since the questionnaire results are tied to the EHR, the clinician can first see the patient's record of long-standing depression and active treatment by the patient's psychiatrist or care team, then this information can be taken into account and collaboration may ensue. Rapid communication among EHR users facilitates improved data completeness and the ability to make timely clinical decisions while reducing administrative burdens and costs to healthcare organizations.

The integration of PROMs into EHR systems also means PROMs data may be harnessed quickly and meaningfully shared between patients and clinicians. For example, native integration of PROMs in the EHR, in contrast to use of PROMs located on third-party sites, minimizes disruption in clinical workflow. The all-in-one integrated platform collects data through the patient portal, tablets, kiosks, or direct entry into the EHR system during patient encounters, which facilitates the creation of reports on PROM data with less hindrance from complexities of third-party PROM tool integration. The capacity for on-demand and trending data visualization for potential sharing with patients has also improved due to native EHR integration and prioritization of ePROs for clinicians. For example, a visualization tool such as Epic Synopsis Report from Epic Systems (Fig. 4.6) depicts both up-to-date and trended PROMs data that reflects a patient's progress and response to treatments.



Months Days All	Aug 2021 8/19/2021	Most Recent Value 11/10/2019 - 11/9/2021	
interfere with ...			
in the past 7 days			
☑ PROMIS Pain Interference T-Score		50	8/13/2021
I could not stop feeling sad.		Sometimes	8/8/2021
I felt alone.		Almost Always	8/8/2021
I felt everything in my life went ...		Never	8/8/2021
I felt like I couldn't do anything ...		Sometimes	8/8/2021
I felt lonely.		Sometimes	8/8/2021
I felt sad.		Almost Always	8/8/2021
I felt unhappy.		Never	8/8/2021
It was hard for me to have fun.		Sometimes	8/8/2021
in the past 7 days			
☑ PROMIS Ped Depressive Symptoms ...		62	8/8/2021
Promis Short Form V2.0<Physical ...			
Promis Short Form V2.0-Physical ...			
Does your health now limit you ...	Cannot do	Cannot do	8/19/2021
Does your health now limit you ...	Cannot do	Cannot do	8/19/2021
Does your health now limit you ...	Cannot do	Cannot do	8/19/2021
Does your health now limit you ...	Cannot do	Cannot do	8/19/2021
Does your health now limit you ...	Cannot do	Cannot do	8/19/2021
Are you able to do chores such ...	Unable to do	Unable to do	8/19/2021
Are you able to dress yourself ...	Unable to do	Unable to do	8/19/2021
Are you able to shampoo	Unable to do	Unable to do	8/19/2021

Fig. 4.6 Trended patient's outcome data in a Synopsis report from Epic Systems (© 2021 Epic Systems Corporation)

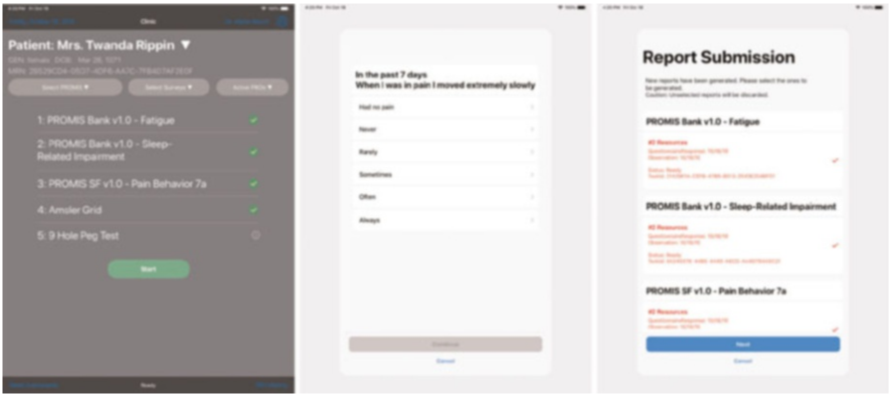
From a patient's perspective, responding to questionnaires before visits means the PROM data can be summarized and viewed immediately prior to seeing providers at the time of appointment [53]. Actual usage of this protocol requires planned and ongoing training for administrators and clinicians to understand not only the interpretation of the scores, but also visualization from the PROMs data to discuss with patients. The data must be communicated in a way that will engage the patient's interest and seem relevant to their context. Otherwise, completing the PROM will likely result in non-use by clinicians and poor rate of data completeness by patients when confronted with other questionnaires in the future.

Another key advancement in EHR systems is the technology that allows the integration of patient-generated health data (PGHD) from smartphones or wearable devices with PROMs. An industry-standard, interoperable SMART (Substitutable Medical Applications and Reusable Technologies) on FHIR (Fast Health Interoperability Resources) technology, can provide additional context to patient's health status such as changes to vital signs or blood sugar level potentially impacting physical function captured in their own devices or apps.

Integration with the EHR allows for synchronized, timely remote monitoring of patient's symptoms and their health status following treatment or, as in a rehabilitation setting, independent home exercise programs prescribed by physical therapists. Perhaps the biggest vantage point for integrating patient-generated data from other devices is that it forces the EHR to solve problems of interoperability, therefore, providing the ability to communicate and exchange pertinent patient data between different health organizations. Moreover, the rapid, scalable data collection also allows management of population health outcomes as data are collected, analyzed, and exchanged with different health systems or entities using the same EHR systems.

Additional benefits from using PROs that are integrated into EHR systems include: assisting patients in remembering their symptoms to report to providers based on questionnaires asked (Fig. 4.7) [143]; monitoring patient's status and responses to interventions and treatments over time [145]; improving patient engagement and shared decision-making by allowing patients to actively control their health data [100]; increasing patients' likelihood of compliance and self-management of their health status [61]; and providing a platform for patients to tell their side of the story using PROMs in clinical settings. The timeliness of data allows clinicians to make relevant decisions based on the latest information and likely to increase patient engagement and data completeness. Patients are more likely to complete the data and stay engaged when clinicians reference and share the outcome data [61, 100].

While the standardization of PROMs and their collection methods have traditionally been a challenge for multiple stakeholders in health systems or private clinics, EHR-related technology has taken a giant step forward in integrating PROs into EHR systems and interoperability. However, EHR systems are not without limitations. Selecting and administering the 'right' PROMs are more difficult for primary care physicians (PCPs) compared to specialists (or allied health professionals such as PTs) due to the wide range of medical conditions PCPs treat.



Patient: Mrs. Twanda Rippin ▼

1: PROMIS Bank v1.0 - Fatigue ✓

2: PROMIS Bank v1.0 - Sleep-Related Impairment ✓

3: PROMIS SF v1.0 - Pain Behavior 7a ✓

4: Amsler Grid ✓

5: 9 Hole Peg Test ○

In the past 7 days
When I was in pain I moved extremely slowly

- Had no pain
- Never
- Rarely
- Sometimes
- Often
- Always

Report Submission

New reports have been generated. Please select the ones to be generated.
Caution: Unselected reports will be discarded.

PROMIS Bank v1.0 - Fatigue ✓

PROMIS Bank v1.0 - Sleep-Related Impairment ✓

PROMIS SF v1.0 - Pain Behavior 7a ✓

Fig. 4.7 Example of Patient-Reported Outcome Measures. (Taken from Sayeed et al. [143])

Adding a measure to the EHR system requires multiple layers of bureaucracy and extensive time even after convincing the vendor to make the change. Choosing the most relevant measure among disparate PROMs is challenging and makes personalized care more difficult [71]. An individual's primary conditions cannot always be assessed in a timely fashion or generalized for comparisons with further data analysis [71].

Currently, PROM incorporation has no cookie cutter template that allows a 'one-size-fits-all' approach to integrating PROs electronically in clinical settings. Their use depends on several factors such as patient context, chief complaints, and psychosocial factors. Perhaps a more concerning matter is the potential inequity of access to technology for certain racial and ethnic groups despite all the technological advancements for enhancing communication between patients and clinicians. Li et al reported that African American and Hispanic patients were least likely to have online patient portal accounts (50%, 57.3%, respectively, compared to 83.9% of white patients) among the rheumatology patient population in a large urban medical center [104]. And while there is no "best of breed" PROM technology due to the ongoing challenges of interoperability and standardization, advancement in the functionality of EHR integration with PROM, and its accessibility, reporting, and analytics of patient data, all have made progress toward person-centered care.

4.4.3 *When Practical Convenience Means Telehealth*

Person-centered care sometimes means that technical complexity and practical convenience must merge: the clinician must take scientifically rigorous measurement and intervention to the patient rather than having the patient come get treatment on the clinician's sophisticated equipment. Electronic platforms for collecting PROMs have helped because the patient with internet access can take PROMs before a clinical visit. Home healthcare remains a viable option for clinicians to gather performance data in the patient's everyday home environment. However, home health still requires a burden of transportation time and resources for the clinician. The growing use of technology to deliver services remotely can improve access and convenience for both the clinician and patient.

Telehealth technology is defined as the use of electronic communications to deliver healthcare services, support, and information remotely to improve patient care [136]. No longer is remote rehabilitation treatment a fringe idea mainly for patients in underserved regions or with limited access to high-quality care. For decades, patients in rural areas relied on medical services from "critical access" hospitals, which are eligible rural hospitals federally funded to improve population health [1]. Patients that required increasingly specialized services had greater difficulty getting it. Underserved populations are further disadvantaged because of hospital closures and decreasing retention of clinicians, including physical therapists.

Telehealth holds potential solutions to address health disparity among those needing specialized services such as neurological or orthopedic rehabilitation. Evidence has shown the efficacy of telehealth when providing physical therapy in different clinical conditions such as low back pain, total joint arthroplasty, and chronic conditions compared to physical therapy in other health settings (e.g., SNFs, home health, inpatient, outpatient) [34, 41, 87, 102]. The advantages for patients are numerous: receiving equitable care even when distance is a barrier for patients (i.e., in rural areas), reducing costs of travel and time, receiving education and training for multiple caregivers or family members, and checking in remotely to monitor critical health conditions.

Patient satisfaction from telehealth shows positive feedback regarding different aspects, including technical setup and overall user experience. For example, Miller et al. used a patient satisfaction survey that contained ten items across multiple domains (e.g., connectivity to telehealth, using telehealth, hearing, seeing, feeling safe, comfort, experience with a physical therapist, meeting expectations, and overall satisfaction). Patients rated their satisfaction for each item on a 5-point Likert scale [116]. For instance, to assess a level of effectiveness of telehealth with regard to the patient's need, a question such as "how did the telehealth physical therapy session (s) meet your needs/expectations" was asked on a 5-point Likert scale (1: Not at all satisfied, 3: Satisfied, 5: Very satisfied) [116].

The authors reported that 94% of 307 patients were satisfied with their telehealth sessions during the early stages of the COVID-19 pandemic [116]. While the target population was primarily from an outpatient rehabilitation environment, specialty areas also included neurologic and pelvic health [116]. And though a vast majority of patients (92%) responded that they were willing to have additional telehealth sessions [116], additional supporting evidence reveals that synchronous or real-time telerehabilitation effectively improves patient outcomes in quality-of-life measures such as physical function and disability, and pain [39]. Further, a hybrid approach to telehealth, combining it with in-person care, is also perceived more favorably by patients than a completely remote episode of care [39].

One of the perceived challenges to rehabilitative assessment when using telehealth is the validity and reliability of remote versions of measures. A systematic review revealed that telerehabilitation assessments demonstrated good concurrent validity for a wide range of outcome measures used in physical therapy practice: pain, gait, balance (e.g., Tinetti test, Berg test), muscle strength, range of motion, functional assessment (e.g., Timed Up and Go), and pain [107]. However, Mani et al. also reported that other special orthopedic tests, neurodynamic tests (e.g., Straight Leg Raises), and postural assessment in the lumbar spine had low to moderate concurrent validity [107].

A potential loss of agreement between telehealth and in-person assessment with respect to clinical management may hinder decision-making in physical therapy. For example, after collecting objective measures and other clinical information about a patient, accurate screening and diagnosis must be reliable regardless what "mode" of treatment is being delivered. Supporting evidence shows a high level of clinical decision agreements between telehealth and in-person assessments among chronic neurological and orthopedic populations [40]. While a high level of agreement (e.g., 83%) in clinical screenings and diagnosis between telehealth and in-person appear favorable, the study had a relatively small sample. A more advanced level of clinical expertise by providers may also be an important factor to deliver consistently high-quality telehealth care.

The rapid adoption rate for telehealth as a response to the COVID-19 pandemic deserves attention. During the early pandemic period, local and state policies were enacted in conjunction with the Centers for Disease Control (CDC) recommendations for healthcare providers to offer a virtual platform including telehealth to slow the spread of the virus that might occur if patients came in for treatment [93]. Surveillance data showed a 154% increase in telehealth visits compared to the previous year during the last week of March 2020, which coincides with the early shelter-in-place orders [93]. A similar study was reported on the surge in telehealth utilization, specifically in an outpatient physical therapy setting at an urban medical center, accounting for 84% of all follow-up visits during early periods of the pandemic [116]. Such evidence of telehealth adoption may indicate that it is a safe and effective platform for patients, clinicians, and the public to minimize exposure to infection. Telehealth also supports the continuity of care from in-person to remote clinical management. For example, the high-risk populations with underlying medical conditions susceptible to infection may find it particularly valuable to continue to rely on rehabilitation and medical care facilitated and accessible by telehealth.

However, telehealth may also exacerbate potential health inequities among racial and ethnic minorities. For example, despite patient-centered, value-added benefits from telehealth, some patients living in lower socioeconomic areas may not own basic smart devices or phones, or have access to stable internet connections. Consequently, a loss of opportunity to receive quality healthcare puts those who may be most vulnerable at risk of further adverse impacts on their quality of life. Even the basic limitations of Internet access due to inadequate infrastructures in their neighborhoods will increase the difficulty of accessing telehealth sessions [113, 135].

One study showed patients living in lower socioeconomic status neighborhoods were significantly less likely to choose telehealth visits [135]. Further, patients with non-English preferences may not choose telehealth given the option of in-person office visit [135]. Those unfamiliar with the use of technology or unable to afford it due to financial constraints will carry the greatest burden of telehealth barriers [135]. In addition, higher acuity level for the person's condition may require an in-person physical examination or diagnostic testing [93], especially when ongoing remote monitoring of critical conditions is time-sensitive.

While telehealth technology continues to be adopted worldwide, the innovation of wearable devices (WD) and phone apps for remote patient monitoring could complement holistic care for patients. Wearable devices can be worn, carried in a pocket, or attached to clothing, to monitor and track health information with minimal to no disruption to patients. The Fitbit is an example of a fitness and wellness promotion device worn as a wrist-watch. Different versions provide wearers with the possibility of setting their own goals for activity, and monitor their progress toward clinically established goals.

This valuable information can be communicated with clinicians or exchanged between health partners or different health organizations using EHR interoperability. The convenient nature of wearing a non-invasive device is that patients can be mobile in their natural environments, whether they are at home or work without significant barriers of physical restrictions using WD. A systematic review reported several clinical indications for WD use with telehealth in chronic populations such as cardiac disease, diabetes, chronic obstructive pulmonary disease [91]. For example, uses might include self-management support for patients, clinical decision support for clinicians, and integration with clinical information systems (i.e. part of EHR systems) in healthcare organizations and outpatient settings [91].

The advancement of digital modalities, wearable devices, and telehealth have facilitated innovative ways to treat different neuromusculoskeletal conditions in rehabilitation. Devices like smart devices, apps, mobile devices and the more affordable wearable devices have pushed the envelope from asynchronous remote monitoring of patients to real-time communication. However, clinicians and healthcare teams still need to consider important implications before prematurely advocating for technology. Person-centered care must always involve the patients early and throughout the process to ensure practical and equitable healthcare with or despite technology.

4.4.4 Case A, Navigating Scientific Rigor and Practical Convenience

The rehabilitation team opens the patient's chart from the electronic health record together with the patient and the team reviews the blood oxygen levels, vital capacity, weight changes, and MAM-CAT scores indicating body structure and function effects of the ALS course in the last month. They detect the trend for the ALS-FRS-R and, based on the literature and information from an ALS data repository (https://nctu.partners.org/MNDS/mnds_als_data_repository), they can project a possible time course for the next few months. They include information from the additional coping and fatigue measures and discuss what the patient and his family deem most important both for the short-term and long-term.

Based on the information shared, the patient and family choose to wait until the next quarterly meeting before starting non-invasive respiratory assistive devices, initially just for better sleep at night. Some re-arrangement of living areas will ultimately become necessary for easier mobility because the patient and family want him to be home for as long as possible. They expect to take advantage of telerehabilitation to ease the burden of transport for some of the upcoming meetings. For the time being, however, they want to optimize mobility and participation in social activities. Thus, they endorse rehabilitation to maximize flexibility and efficiency of movement and recommend assistive devices to increase function with less energy expenditure. The patient and family express their appreciation for the team's emphasis on meaningful outcomes and the patient's preferences throughout the experience.

4.5 Summary and Recommendations

This section summarizes both the challenges and promises for improving clinical practice through person-centered outcome measurement. Although the dilemmas facing measurers present real challenges when choosing clinically relevant measures, the ideal of brilliant results for every person should motivate stakeholders to continue striving. Success will mean that consistently high quality and equitable healthcare tailored to each specific person's unique situation becomes the norm.

However, no one person or stakeholder role can meet the challenges alone. Many individuals will need to participate. The first section of this chapter discussed the contributions of both the clinician and environment to measurement choices. The clinician and healthcare environment must both evolve. The clinician must move past outdated assumptions about measuring; the healthcare environment must effectively emphasize health equity with the diversity of measures and experiences required in person-centered practice. The healthcare environment shows some promise: it has progressed toward valuing patient-reported outcomes with policy support from the legislature and processes in place through PROMIS and PCORI. Patients now participate in planning and funding research focused on person-centered outcomes. The dissemination of this type of participation throughout healthcare may help bridge the gap between adopters and the rest of society.

The second section discussed personalization and standardization and their potential association with equality and equity. Patient groupings along meaningful categories can help direct care more appropriately for both efficiency and effectiveness. The section also presented theoretical constructs such as the ICF and systems frameworks that can help stakeholders keep standard measurement personalized by choosing measures according to the person and environmental context in which they live and work.

The third section discussed satisfaction and effectiveness. While assessing the person's experience in healthcare through client-centered measures such as the CCRQ can inform practice, satisfaction with the experience does not substitute for measures of effectiveness. Multiple types of outcome measures can help determine if healthcare effectively meets the person's needs, including health and disease biomarkers and wearables, attention to nonlinear variability within time-sequenced data, disability as the gap between having and valuing, the use of QALYs and DALYS, and assessing both current and preferred abilities. Effectiveness measures may also include assessing adverse conditions that an intervention may minimize, mediators of patient's response to treatment, and health-related quality of life.

The fourth section discussed scientific rigor and practical convenience, both of which require extensive expertise and diligence to get technically sound measures practically available, so clinicians have all they need at their fingertips. This requires collaboration at multiple levels. Best practice in measurement development calls for patients and clinicians to work together to define constructs and the scales on which they will be measured. Psychometricians can ensure that resulting measures meet the standards for scientific rigor. Researchers must contribute comparative studies to support recommendations of measures for common constructs. Computer experts may program a computer-adaptive test or provide a platform for access to the measures of interest or a data repository. Healthcare administrators must navigate a budget for servers and electronic health record vendors. Vendors must attend to the needs of the clinicians who demand that their preferred measure gets supported while simultaneously optimizing profitability for shareholders. Policymakers must continue to strive to make systems interoperable so that patients can move through a course of care or between providers and carry their excellent healthcare along with them.

These discussions lead to four recommendations for improving practice.

1. First, collaborations among content specialists and psychometricians must support measure development and modification; they must ensure that rehabilitation teams have measures for both qualitative and quantitative data that have a theoretical and person-centered foundation.
2. Second, scientifically rigorous measures must also meet criteria for accessibility and convenience, which means that developers must collaborate with computer programmers and platform creators who then test ease of use by patients.
3. Third, data managers and vendors must work together to build and support systems that are both flexible and durable while also secure and easy for patients and clinicians to view.
4. Fourth, administrators will need both satisfaction and effectiveness data to demonstrate clinical practice quality and improvement over time.

The ideal team, then, would comprise the care team and patient, IT support, domain experts, psychometricians, and third-party vendors associated with electronic health records. The objective is to leverage outcome measures to optimize equitable and person-centered healthcare and improve clinical practice. Every link in the chain must be person-centered if brilliant processes can be used to achieve brilliant results when used by everyday people.

Are we there, yet? No. But the evidence shows areas where we have gotten closer. The next steps in the process involve collaborations at all levels. Only then will we improve clinical practice with brilliant person-centered outcome measurement.

References

1. 2019-07 | CMS. (n.d.). Retrieved January 9, 2021, from https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/MLN-Publications-Items/CMS1243364
2. M. Abdul-Kareem. *Here's why we should care more about equity, not equality* (2018, January 5). Retrieved January 9, 2022, from https://muslimgirl.com/heres-care-equity-equality/
3. A. Alismail, B. Schaeffer, A. Oh, S. Hamiduzzaman, N. Daher, H.Y. Song, et al., The use of the Net Promoter Score (NPS) in an outpatient allergy and pulmonary clinic: an innovative look into using tablet-based tool vs traditional survey method. Patient Relat. Outcome Meas. **11**, 137–142 (2020). https://doi.org/10.2147/PROM.S248431
4. D.D. Allen, Proposing 6 dimensions within the construct of movement in the Movement Continuum Theory. Phys. Ther. **87**, 888–898 (2007a)
5. D.D. Allen, Responsiveness of the movement ability measure: A self-report instrument proposed for assessing the effectiveness of physical therapy intervention. Phys. Ther. **87**, 917–924 (2007b)
6. D.D. Allen, Validity and reliability of the movement ability measure: A self-report instrument proposed for assessing movement across diagnoses and ability levels. Phys. Ther. **87**, 899–916 (2007c)
7. D.D. Allen, Using item response modeling methods to test theory related to human performance. J. Appl. Measurement **11**(2), 99–111 (2010)
8. D.D. Allen, Neuromuscular Diseases, in *Umphred's Neurological Rehabilitation*, ed. by R. T. Lazaro, 7th edn., (Elsevier, Maryland Heights, 2019)
9. D.D. Allen, C.A. Cott, Evaluating rehabilitation outcomes from the client's perspective by identifying the gap between current and preferred movement ability. Disabil. Rehabil. **32**(6), 452–461 (2010)
10. D.D. Allen, J.M. Wagner, Assessing the gap between current movement ability and preferred movement ability as a measure of disability. Phys. Ther. **91**, 1789–1803 (2011). https://doi.org/10.2522/ptj.20100393
11. D.D. Allen, P. Ni, S.M. Haley, Efficiency and sensitivity of multidimensional computerized adaptive testing of pediatric physical functioning. Disabil. Rehabil. **30**(6), 479–484 (2008)
12. D.D. Allen, C. Talavera, S. Baxter, K. Topp, Gaps between patients' reported current and preferred abilities versus clinicians' emphases during an episode of care: Any agreement? Qual. Life Res. **24**(5), 1137–1143 (2015). https://doi.org/10.1007/s11136-014-0888-0
13. D.D. Allen, C. Talavera, S. Baxter, K. Topp. *Patient versus Clinician focus of care: Who's minding the gap?* [abstract]. Paper presented at the IV STEP Conference, American Physical Therapy Association, Columbus, OH, 2016. Poster presentation retrieved from https://u.osu.edu/ivstep/poster/abstracts/001-allen-et-al/

14. F. Augustovski, L.D. Colantonio, J. Galante, A. Bardach, J.E. Caporale, V. Zárate, et al., Measuring the benefits of healthcare: DALYs and QALYs – Does the choice of measure matter? A case study of two preventive interventions. Int. J. Health Policy Management **7**(2), 120–136 (2018). https://doi.org/10.15171/ijhpm.2017.47
15. C.C. Austin, S. Brown, N. Fong, C. Humphrey, A. Leahey, P. Webster, Research data repositories: Review of current features, gap analysis, and recommendations for minimum requirements. IASSIST Q **39**(4), 24 (2016). https://doi.org/10.29173/iq904
16. T. Baranowski, D.D. Allen, L. Mâsse, M. Wilson, Does participation in an intervention affect responses on self-report questionnaires? Health Educ. Res. **21**(Supplement 1), i98–i109 (2006)
17. E. Basch, Patient-reported outcomes—Harnessing patients' voices to improve clinical care. N. Engl. J. Med. **376**, 105–108 (2017). https://doi.org/10.1056/NEJMp1611252
18. J.F. Baumhauer, Patient-reported outcomes—Are they living up to their potential? New Engl. J. Med. **377**(1), 6–9 (2017). https://doi.org/10.1056/nejmp1702978
19. M. Beninato, L.G. Portney, Applying concepts of responsiveness to patient management in neurologic physical therapy. J. Neurol. Phys. Ther. **35**, 75–81 (2011)
20. D. Berghuis-Kelley, S. Scherer, Outcome measures in cardiopulmonary physical therapy: Use of the Patient Specific Functional Scale. Cardiopulm. Phys. Ther. J. **18**(3), 21–23 (2007)
21. M. Bergner, R.A. Bobbitt, W.B. Carter, B.S. Gilson, The sickness impact profile: Development and final revision of a health status measure. Med. Care **19**, 787–805 (1981)
22. V.A.J. Block, E. Pitsch, P. Tahir, B.A.C. Cree, D.D. Allen, J.M. Gelfand, Remote physical activity monitoring in neurological disease: A systematic review. PloS One **11**(4), e0154335 (2016). https://doi.org/10.1371/journal.pone.0154335
23. V.J. Block, A. Lizee, E. Crabtree-Hartman, C.J. Bevan, J.S. Graves, R. Bove, et al., Continuous daily assessment of multiple sclerosis disability using remote step count monitoring. J. Neurol. **264**, 316–326 (2017)
24. V.J. Block, R. Bove, C. Zhao, P. Garcha, J. Graves, A.R. Romeo, et al., Association of continuous assessment of step count by remote monitoring with disability progression among adults with multiple sclerosis. JAMA Netw. Open **2**(3), e190570 (2019). https://doi.org/10.1001/jamanetworkopen.2019.0570
25. J.M. Bollinger, P.D. Zuk, M.A. Majumder, E. Versalovic, A.G. Villanueva, R.L. Hsu, et al., What is a medical information commons? J. Law Med. Ethics **47**, 41–50 (2019). https://doi.org/10.1177/1073110519840483
26. G. Borg, *Borg's Perceived Exertion and Pain Scales* (Human Kinetics, Champaign, 1998)
27. E.H. Bradford, I. Baert, M. Finlayson, P. Feys, J. Wagner, Feasibility of an international multiple sclerosis rehabilitation data repository: Perceived challenges and motivators for sharing data. Int. J. MS Care **20**(1), 17–26 (2018). https://doi.org/10.7224/1537-2073.2016-009
28. N. Breen, D. Berrigan, J.S. Jackson, D.W.S. Wong, F.B. Wood, J.C. Denny, et al., Translational health disparities research in a data-rich world. Health Equity **3**, 588–600 (2019). https://doi.org/10.1089/heq.2019.0042
29. J. Burridge, M.A. Murphy, J. Buurke, P. Feys, T. Keller, V. Klamroth-Marganska, et al., A systematic review of international clinical guidelines for rehabilitation of people with neurological conditions: What recommendations are made for upper limb assessment? Front Neurol. **10**(567) (2019). https://doi.org/10.3389/fneur.2019.00567
30. K.S. Button, J.P.A. Ioannidis, C. Mokrysz, B.A. Nosek, J. Flint, E.S.J. Robinson, M.R. Munafò, Power failure: Why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. **14**(5), 365–376 (2013). https://doi.org/10.1038/nrn3475
31. J.T. Capell, T. Alexander, J. Pryor, M. Fisher, Patient reported experience of inpatient rehabilitation in Australia. Patient Exp. J. **7**(3) (2020). https://doi.org/10.35680/2372-0247.1424
32. J.M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, et al., The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS Study Group (Phase III). J. Neurol. Sci. **169**, 13–21 (1999)

33. L. Chaker, A. Falla, S.J. van der Lee, T. Muka, D. Imo, L. Jaspers, et al., The global impact of non-communicable diseases on macro-economic productivity: A systematic review. Eur. J. Epidemiol. **30**(5), 357–395 (2015). https://doi.org/10.1007/s10654-015-0026-5
34. C.L. Christiansen, M.J. Miller, A.M. Murray, R.O. Stephenson, J.E. Stevens-Lapsley, W.R. Hiatt, M.L. Schenkman, Behavior-change intervention targeting physical function, walking, and disability after dysvascular amputation: A randomized controlled pilot trial. Arch. Phys. Med. Rehabil. **99**(11), 2160–2167 (2018). https://doi.org/10.1016/j.apmr.2018.04.011
35. E.T. Cohen, K. Potter, D.D. Allen, G.L. Widener, A.M. Yorke, S. Bennett, K.G. Brandfass, Selecting rehabilitation outcome measures for persons with multiple sclerosis. Int. J. MS Care **17**, 181–189 (2015)
36. C.A. Cott, Client-centred rehabilitation: Client perspectives. Disabil. Rehabil. **26**(24), 1411–1422 (2004)
37. C.A. Cott, E. Finch, D. Gasner, K. Yoshida, S.G. Thomas, M.C.M. Verrier, The movement continuum theory of physical therapy. Physiother. Canada **47**(2), 87–95 (1995)
38. C.A. Cott, G. Teare, K.S. McGilton, S. Lineker, Reliability and construct validity of the client-centred rehabilitation questionnaire. Disabil. Rehabil. **28**(22), 1387–1397 (2006)
39. M.A. Cottrell, O.A. Galea, S.P. O'Leary, A.J. Hill, T.G. Russell, Real-time telerehabilitation for the treatment of musculoskeletal conditions is effective and comparable to standard practice: A systematic review and meta-analysis. Clin. Rehabil. **31**, 625–638 (2017). https://doi.org/10.1177/0269215516645148
40. M.A. Cottrell, S.P. O'Leary, P. Swete-Kelly, B. Elwell, S. Hess, M.A. Litchfield, et al., Agreement between telehealth and in-person assessment of patients with chronic musculoskeletal conditions presenting to an advanced-practice physiotherapy screening clinic. Musculoskelet. Sci. Pract. **38**, 99–105 (2018). https://doi.org/10.1016/j.msksp.2018.09.014
41. A.B. Dario, A. Moreti Cabral, L. Almeida, M.L. Ferreira, K. Refshauge, M. Simic, et al., Effectiveness of telehealth-based interventions in the management of non-specific low back pain: A systematic review with meta-analysis. Spine J. **17**, 1342–1351 (2017). https://doi.org/10.1016/j.spinee.2017.04.008
42. Data Repository | NNLM. (n.d.). Retrieved January 1, 2021, from https://nnlm.gov/data/thesaurus/data-repository
43. J. Daubenmier, M.T. Chao, W. Hartogensis, R. Liu, P.J. Moran, M.C. Acree, et al., Exploratory analysis of racial/ethnic and educational differences in a randomized controlled trial of a mindfulness-based weight loss intervention. Psychosom. Med. **83**(6), 503–514 (2020). https://doi.org/10.1097/PSY.0000000000000859
44. R.A. Deyo, S.F. Dworkin, D. Amtmann, G. Andersson, D. Borenstein, E. Carragee, et al., Report of the NIH task force on research standards for chronic low back pain. J. Pain **15**(6), 569–585 (2014). https://doi.org/10.1016/j.jpain.2014.03.005
45. C. Donado, K. Lobo, C.B. Berde, F.T. Bourgeois, Developing a pediatric pain data repository. JAMIA Open **3**(1), 31–36 (2020). https://doi.org/10.1093/jamiaopen/ooz062
46. J. Dorst, A.C. Ludolph. Non-invasive ventilation in amyotrophic lateral sclerosis. Ther. Adv. Neurol. Disord. (2019). https://doi.org/10.1177/1756286419857040
47. M. Dougados, A. Moore, S. Yu, X. Gitton, Evaluation of the patient acceptable symptom state in a pooled analysis of two multicentre, randomised, double-blind, placebo-controlled studies evaluating lumiracoxib and celecoxib in patients with osteoarthritis. Arthritis. Res. Ther. **9**(R11) (2007). https://doi.org/10.1186/ar2118
48. S. Dremann. A vibrant life, *Palo Alto Weekly* (2018, July 27), pp. 12–16.
49. T. Ewert, D.D. Allen, M. Wilson, B. Ustun, G. Stucki, Validation of the international classification of functioning disability and health framework using multidimensional item response modeling. Disabil. Rehabil. **32**(17), 1397–1405 (2010)
50. J.C. Fairbank, J. Couper, J.B. Davies, J.P. O'Brien, The Oswestry low back pain disability questionnaire. Physiotherapy **66**(8), 271–273 (1980)

51. M. Fisher, J. Pryor, J. Capell, T. Alexander, F. Simmonds, The psychometric properties of a modified client-centred rehabilitation questionnaire in an Australian population. Disabil. Rehabil. Assist. Technol. **42**(1), 122–129 (2020). https://doi.org/10.1080/09638288.2018.1494214
52. J.D. Fisk, P.G. Ritvo, L. Ross, D.A. Haase, T.J. Marrie, W.F. Schlech, Measuring the functional impact of fatigue: Initial validation of the fatigue impact scale. Clin. Infect. Dis. **18**(Suppl 1), S79–S83 (1994)
53. H.H. Forsberg, E.C. Nelson, R. Reid, D. Grossman, M.P. Mastanduno, L.T. Weiss, et al., Using patient-reported outcomes in routine practice: Three novel use cases and Implications. J. Ambul. Care Management **38**(2), 188–195 (2015). https://doi.org/10.1097/JAC.0000000000000052
54. J.K. Freburger, G.M. Holmes, L.J. Ku, M.P. Cutchin, K. Heatwole-Shank, L.J. Edwards, Disparities in postacute rehabilitation care for stroke: An analysis of the state inpatient databases. Arch Phys Med Rehabil **92**(8), 1220–1229 (2011). https://doi.org/10.1016/j.apmr.2011.03.019
55. S. Fritz, M. Lusardi, White Paper: "Walking Speed: the Sixth Vital Sign". J. Geriatric Physical Ther. **32**(2), 2–5 (2009)
56. A.W. Garcia, A.C. King, Predicting long-term adherence to aerobic exercise: A comparison of two models. J. Sport Exercise Psychol. **13**, 394–410 (1991)
57. GBD 2017 DALYs and HALE Collaborators. (2018). Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. Erratum in: Lancet. 2019 Jun 22;393(10190):e44. . Lancet, 392(10159), 1859–1922. https://doi.org/10.1016/S0140-6736(18)32335-3
58. M. Gerard, A. Fossa, P.H. Folcarelli, J. Walker, S.K. Bell, What patients value about reading visit notes: a qualitative inquiry of patient experiences with their health information. J. Med. Internet Res. **19**(7), e237 (2017). https://doi.org/10.2196/jmir.7212
59. C. Gibbons, F. Pagnini, T. Friede, & Young, C. A. (2018). Treatment of fatigue in amyotrophic lateral sclerosis/ motor neuron disease. Cochrane Database of Systematic Reviews (CD011005).
60. R.E. Gliklich, N.A. Dreyer, & M.B. Leavy. *Patient Registries* (2014). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK208643/
61. H.T. Gold, R.J. Karia, A. Link, R. Lebwohl, J.D. Zuckerman, T.J. Errico, et al., Implementation and early adaptation of patient-reported outcome measures into an electronic health record: A technical report. Health Infor. J. **26**(1), 129–140 (2020). https://doi.org/10.1177/1460458218813710
62. S.M. Haley, M.A. Fragala-Pinkham, Interpreting change scores of tests and measures used in physical therapy. Phys. Ther. **86**(5), 735–743 (2006)
63. S.M. Haley, W.J. Coster, L.H. Ludlow, J.T. Haltiwanger, P.A. Andrellos, *Pediatric evaluation of disability inventory: Development, standardization and administration manual* (Trustees of Boston University, Boston, MA, 1992)
64. S.M. Haley, P.L. Andres, W.J. Coster, M. Kosinski, P.S. Ni, A. Jette, Short-form activity measure for post-acute care (AM-PAC). Arch. Phys. Med. Rehabil. **85**, 649–660 (2004a)
65. S.M. Haley, W.J. Coster, P.L. Andres, M. Kosinski, P.S. Ni, Score comparability of short-forms and computerized adaptive testing: simulation study with the Activity Measure for Post-Acute Care (AM-PAC). Arch. Phys. Med. Rehabil. **85**, 661–666 (2004b)
66. C. Hall, D.D. Allen, Role of a movement system approach in physical therapy management of musculoskeletal pain, in *Physical Therapy Approaches to the Lower Quarter: Independent Study Course 29.1*, ed. by C. Hughes, vol. 2, (Academy of Orthopaedic Physical Therapy, La Crosse, 2019), pp. 35–68
67. R.T. Harbourne, N. Stergiou, Movement variability and the use of nonlinear tools: Principles to guide physical therapist practice. Phys. Ther. **89**, 267–282 (2009)

68. T. Hart, M.P. Dijkers, J. Whyte, L.S. Turkstra, J.M. Zanca, A. Packel, et al., A theory-driven system for the specification of rehabilitation treatments. Arch. Phys. Med. Rehabil. **100**, 172–180 (2019). https://doi.org/10.1016/j.apmr.2018.09.109
69. HHS, & CMS. *2020 Annual Call for Quality Measures Fact Sheet What is the Quality Payment Program?* (2020). Retrieved from https://qpp.cms.gov
70. J. Hobart, D. Lamping, R. Fitzpatrick, A. Riazi, A. Thompson, The multiple sclerosis impact scale (MSIS-29): A new patient-based outcome measure. Brain **124**(Part 5), 962–973 (2001)
71. C.J. Hsiao, C. Dymek, B. Kim, B. Russell, Advancing the use of patient-reported outcomes in practice: Understanding challenges, opportunities, and the potential of health information technology. Qual. Life Res. **28**(6), 1575–1583 (2019). https://doi.org/10.1007/s11136-019-02112-0
72. S. Hudgens, R. Schüler, J. Stokes, S. Eremenco, E. Hunsche, T.P. Leist, Development and validation of the FSIQ-RMS: A new patient-reported questionnaire to assess symptoms and impacts of fatigue in relapsing multiple sclerosis. Value Health **22**(4), 453–466 (2019). https://doi.org/10.1016/j.jval.2018.11.007
73. J.M. Huisinga, J.M. Yentes, M.L. Filipi, N. Stergiou, Postural control strategy during standing is altered in patients with multiple sclerosis. Neurosci. Lett. **524**, 124–128 (2012)
74. J.M. Huisinga, M. Mancini, R.J. St. George, F.B. Horak, Accelerometry reveals differences in gait variability between patients with multiple sclerosis and healthy controls. Ann. Biomed. Eng. **41**, 1670–1679 (2013)
75. C.M. Hunt, G.L. Widener, D.D. Allen, Variability in postural control with and without balance-based torso-weighting in people with multiple sclerosis and healthy controls. Phys. Ther. **94**, 1489–1498 (2014)
76. J. Hurn, I. Kneebone, M. Cropley, Goal setting as an outcome measure: A systematic review. Clin. Rehabil. **20**(9), 756–772 (2006)
77. J.M. Hush, K. Cameron, M. Mackey, Patient satisfaction with musculoskeletal physical therapy care: A systematic review. Phys. Ther. **91**, 25–36 (2011)
78. L.I. Iezzoni, V.A. Freedman, Turning the disability tide: the importance of definitions. J. Am. Med. Assoc. **299**(3), 332–334 (2008)
79. International Classification of Functioning, Disability and Health: ICF. (2001). World Health Organization, Geneva.
80. H. Israel, R.R. Richter, A guide to understanding meta-analysis. J. Orthop. Sports Phys. Ther. **41**, 496–504 (2011)
81. G.P. Jacobson, C.W. Newman, The development of the dizziness handicap inventory. Arch. Otolaryngol. Head Neck Surg. **116**, 424 (1990)
82. A.M. Jette, Opening the Black Box of rehabilitation interventions. Phys. Ther. **100**(6), 883–884 (2020). https://doi.org/10.1093/ptj/pzaa078
83. A.M. Jette, S.M. Haley, W.J. Coster, J.T. Kooyoomijian, S. Levenson, T. Heeren, J. Ashba, Late life function and disability instrument: I. Development and evaluation of the disability component. J. Gerontol. Med. Sci. **57A**(4), M209–M216 (2002)
84. D.U. Jette, J. Halbert, C. Iverson, E. Miceli, P. Shah, Use of standardized outcome measures in physical therapist practice: Perceptions and applications. Phys. Ther. **89**(2), 125–135 (2009). https://doi.org/10.2522/ptj.20080234
85. D.U. Jette, M. Stilphen, V.K. Ranganathan, S.D. Passek, F.S. Frost, A.M. Jette, Validity of the AM-PAC "6-Clicks" inpatient daily activity and basic mobility short forms. Phys. Ther. **94**(3), 379–391 (2014). https://doi.org/10.2522/ptj.20130199
86. D.V. Jewell, *Guide to Evidence-Based Physical Therapy Practice* (Jones and Bartlett, Sudbury, 2008)
87. S. Jiang, J. Xiang, X. Gao, K. Guo, B. Liu, The comparison of telerehabilitation and face-to-face rehabilitation after total knee arthroplasty: A systematic review and meta-analysis. J. Telemed. Telecare **24**(4), 257–262 (2018). https://doi.org/10.1177/1357633X16686748
88. B.M. Joyce, K.J. Rockwood, C.C. Mate-Kole, Use of goal attainment scaling in brain injury in a rehabilitation hospital. Am. J. Phys. Med. Rehabil. **73**, 10–14 (1994)

89. J.H. Kahn, R. Tappan, C.P. Newman, P. Palma, W. Romney, E.T. Stultz, et al., Outcome measure recommendations from the spinal cord injury EDGE task force. Phys. Ther. **96**(11), 1832–1842 (2016). https://doi.org/10.2522/ptj.20150453
90. J.P. Kaipust, J.M. Huisinga, M. Filipi, N. Stergiou, Gait variability measures reveal differences between multiple sclerosis patients and healthy controls. Motor Control **16**, 229–244 (2012)
91. T. Kamei, T. Kanamori, Y. Yamamoto, S. Edirippulige, The use of wearable devices in chronic disease management to enhance adherence and improve telehealth outcomes: A systematic review and meta-analysis. J. Telemed. Telecare (2020). https://doi.org/10.1177/1357633X20937573
92. T. Kiresuk, R. Sherman, Goal attainment scaling: A general method of evaluating comprehensive mental programs. Community Mental Health J. **4**, 443–453 (1968)
93. L.M. Koonin, B. Hoots, C.A. Tsang, Z. Leroy, K. Farris, B. Jolly, et al., Trends in the use of telehealth during the emergence of the COVID-19 pandemic — United States, January–March 2020. MMWR **69**(43), 1595–1599 (2020). https://doi.org/10.15585/mmwr.mm6943a3
94. S. Körner, K. Kollewe, S. Abdulla, A. Zapf, R. Dengler, S. Petri, Interaction of physical function, quality of life and depression in Amyotrophic lateral sclerosis: Characterization of a large patient cohort. BMC Neurol. **15**, 84 (2015)
95. M. Körner, H. Dangel, A. Plewnia, J. Haller, M.A. Wirtz, Psychometric evaluation of the Client-Centered Rehabilitation Questionnaire (CCRQ) in a large sample of German rehabilitation patients. Clin. Rehabil. **31**, 926–935 (2017). https://doi.org/10.1177/0269215516665158
96. D. Kos, G. Nagels, M.B. D’Hooghe, M. Duportail, E. Kerckhofs, A rapid screening tool for fatigue impact in multiple sclerosis. BMC Neurol. **6**(27) (2006). https://doi.org/10.1186/1471-2377-6-27
97. L.B. Krupp, N.G. LaRocca, J. Muir-Nash, A.D. Steinberg, The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. Arch. Neurol. **46**(10), 1121–1123 (1989)
98. J. Kurtzke, Rating neurological impairment in multiple sclerosis: An expanded disability status scale (EDSS). Neurology **33**, 1444–1452 (1983)
99. R.D. Larson, Psychometric properties of the modified fatigue impact scale. Int. J. MS Care **15**, 15–20 (2013)
100. D.C. Lavallee, K.E. Chenok, R.M. Love, C. Petersen, E. Holve, C.D. Segal, P.D. Franklin, Incorporating patient-reported outcomes into health care to engage patients and enhance care. Health Aff. **35**(4), 575–582 (2016). https://doi.org/10.1377/hlthaff.2015.1362
101. M. Law, S. Baptiste, M. McColl, A. Opzoomer, H. Polatajko, N. Pollock, The Canadian occupational performance measure: An outcome measure for occupational therapy. Can. J. Occup. Ther. **57**(2), 82–87 (1990)
102. A.C.W. Lee, M. Billings, Telehealth implementation in a skilled nursing facility: Case report for physical therapist practice in Washington. Phys. Ther. **96**(2), 252–259 (2016). https://doi.org/10.2522/ptj.20150079
103. A. Lerdal, A. Kottorp, Psychometric properties of the fatigue severity scale–Rasch analyses of individual responses in a Norwegian stroke cohort. Int. J. Nurs. Stud. **48**(10), 1258–1265 (2011). https://doi.org/10.1016/j.ijnurstu.2011.02.019
104. J. Li, J. Yazdany, L. Trupin, Z. Izadi, M. Gianfrancesco, S. Goglin, G. Schmajuk, Capturing a patient-reported measure of physical function through an online electronic health record patient portal in an ambulatory clinic: Implementation study. J. Med. Internet Res. **20**(5) (2018). https://doi.org/10.2196/medinform.8687
105. J.F. Malec, J.S. Smigielski, R.W. Depompolo, Goal attainment and outcome measurement in postacute brain injury rehabilitation. Arch. Phys. Med. Rehabil. **72**, 138–143 (1991)
106. E.C. Manchanda, C. Couillard, K. Sivashanker, Inequity in crisis standards of care. New Engl. J. Med. **383**, e16 (2020). https://doi.org/10.1056/NEJMp2011359

107. S. Mani, S. Sharma, B. Omar, A. Paungmali, L. Joseph, Validity and reliability of Internet-based physiotherapy assessment for musculoskeletal disorders: A systematic review. J. Telemed. Telecare **23**(3), 379–391 (2017). https://doi.org/10.1177/1357633X16642369
108. L.C. Mâsse, D.D. Allen, M. Wilson, G.C. Williams, Introducing equating methodologies to compare test scores from two different self-regulation scales. Health Educ. Res. **21**-(Supplement 1), i110–i120 (2006)
109. L. McCluskey, D. Casarett, A. Siderowf, Breaking the news: A survey of ALS patients and their caregivers. Amyotroph. Later. Scler. Other Motor Neuron Disord. **5**, 131–135 (2004)
110. K.L. McCulloch, A.L. de Joya, K. Hays, E. Donnelly, T.K. Johnson, C.D. Nirider, et al., Outcome measures for persons with moderate to severe traumatic brain injury: Recommendations from the American physical therapy association academy of neurologic physical therapy TBI EDGE Task Force. J. Neurol. Phys. Ther. **40**(4), 269–280 (2016). https://doi.org/10.1097/NPT.0000000000000145
111. T. McGinn, P. Wyer, J. Wisnivesky, P.J. Devereaux, I. Stiell, S. Richardson, et al., Advanced topics in diagnosis: Clinical prediction rules, in *Users' Guides to Medical Literature. A Manual for Evidence-Based Clinical Practice*, ed. by G. Guyatt, D. Rennie, M. Meade, D. Cook, 2nd edn., (McGraw Hill Medical, 2008), pp. 491–505
112. C.A. McHorney, J.E. Ware, R. Lu, C.D. Sherbourne, The MOS 36-item short-form health survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med. Care. **32**(1), 40–66 (1994)
113. Measuring digital development: Facts and figures 2020. (n.d.). Retrieved January 8, 2021, from https://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx
114. L.N. Medford-Davis, G.C. Fonarow, D.L. Bhatt, H. Xu, E.E. Smith, R. Suter, et al., Impact of insurance status on outcomes and use of rehabilitation services in acute ischemic stroke: findings from Get With The Guidelines-Stroke. J. Am. Heart. Assoc. **5**(11), e004282 (2016). https://doi.org/10.1161/JAHA.116.004282
115. A. Middleton, G.K. Fitzgerald, A. Delitto, R.B. Saper, K. Gergen Barnett, J. Stevans, Implementing stratified care for acute low back pain in primary care using the STarT Back instrument: A process evaluation within the context of a large pragmatic cluster randomized trial. BMC Musculoskelet. Disord. **21**(1), 776 (2020). https://doi.org/10.1186/s12891-020-03800-6
116. M.J. Miller, S.S. Pak, D.R. Keller, D.E. Barnes, Evaluation of pragmatic telehealth physical therapy implementation during the COVID-19 pandemic. Phys. Ther. (2020). https://doi.org/10.1093/ptj/pzaa193
117. R. Mills, C. Young, R. Nicholas, J. Pallant, A. Tennant, Rasch analysis of the fatigue severity scale in multiple sclerosis. Mult. Scler. **15**(1), 81–87 (2009). https://doi.org/10.1177/1352458508096215
118. R.J. Mills, C.A. Young, J.F. Pallant, A. Tennant, Rasch analysis of the modified fatigue impact scale (MFIS) in multiple sclerosis. J. Neurol. Neurosurg. Psychiatry **81**(9), 1049–1051 (2010)
119. S. Mitra, The capability approach and disability. J. Disabil. Policy Stud. **16**, 236–247 (2006)
120. J.L. Moore, K. Potter, K. Blankshain, S.L. Kaplan, L.C. O'Dwyer, J.E. Sullivan, A core set of outcome measures for adults with neurologic conditions undergoing rehabilitation. J. Neurol. Phys. Ther. **42**, 174–220 (2018)
121. S.Z. Nagi, Disability concepts revisited: Implications for prevention, in *Disability in America: Toward a National Agenda for Prevention*, ed. by A. M. Pope, A. R. Tarlov, (Institute of Medicine, National Academy Press, Washington, DC, 1991)
122. M.G. Ory, P.J. Jordan, T. Bazzarre, The behavior change consortium: Setting the stage for a new century of health behavior-change research. Health Educ. Res. **17**(5), 500–511 (2002)
123. L. Paul, S. Coote, J. Crosbie, D. Dixon, L. Hale, E. Holloway, et al., Core outcome measures for exercise studies in people with multiple sclerosis: Recommendations from a multidisciplinary consensus meeting. Mult. Scler. J. **20**(12), 1641–1650 (2014). https://doi.org/10.1177/1352458514526944

124. E.R. Pfoh, A. Hamilton, B. Hu, M. Stilphen, M.B. Rothberg, The six-clicks mobility measure: A useful tool for predicting discharge disposition. Arch. Phys. Med. Rehabil. **101**(7), 1199–1203 (2020). https://doi.org/10.1016/j.apmr.2020.02.016
125. Physical Therapy Outcomes Registry | PTOR. (n.d.). Retrieved January 4, 2021, from https://www.ptoutcomes.com/
126. S.M. Pincus, Approximate entropy as a measure of system complexity. Proc. Natl. Acad. Sci. U. S. A. **88**(6), 2297–2301 (1991)
127. S.M. Pincus, I.M. Gladstone, R.A. Ehrenkranz, A regularity statistic for medical data analysis. J. Clin. Monit. **7**(4), 335–345 (1991)
128. K. Potter, G. Fulk, Y. Salem, J. Sullivan, Outcome measures in neurological physical therapy practice: Part I. Making sound decisions. J. Neurol. Phys. Ther. **35**, 57–64 (2011)
129. K. Potter, E.T. Cohen, D.D. Allen, S.E. Bennett, K.G. Brandfass, G.L. Widener, A.M. Yorke, Outcome measures for individuals with multiple sclerosis: Recommendations from the American Physical Therapy Association Neurology Section Task Force. Phys. Ther. **94**, 593–608 (2014)
130. L.E. Powell, A.M. Myers, The activities-specific balance confidence (ABC) scale. J. Gerontol. **50A**(1), M28–M34 (1995)
131. J. Pryor, M. Fisher, An examination of the CCRQ as a measure of person-centred rehabilitation. J. Australas. Rehabil. Nurses Assoc. **22**, 22–26 (2019). https://doi.org/10.33235/jarna.22.2.22-26.2019
132. G.E.A. Pust, J. Pöttgen, J. Randerath, S. Lau, C. Heesen, S.M. Gold, I.-K. Penner, In search of distinct MS-related fatigue subtypes: Results from a multi-cohort analysis in 1.403 MS patients. J. Neurol. **266**, 1663–1673 (2019). https://doi.org/10.1007/s00415-019-09311-2
133. A.E. Raczek, J.E. Ware, J.B. Bjorner, B. Gandek, S.M. Haley, N.K. Aaronson, et al., Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: Results from the IQOLA Project (International Quality of Life Assessment). J. Clin. Epidemiol. **51**(11), 1203–1214 (1998). https://doi.org/10.1016/s0895-4356(98)00112-7
134. P. Räsänen, E. Roine, H. Sintonen, V. Semberg-konttinen, O. Ryynänen, R. Roine, Use of quality-adjusted life years for the estimation of effectiveness of health care: A systematic literature review. Int. J. Technol. Assess. Health Care **22**(2), 235–241 (2006)
135. M.E. Reed, J. Huang, I. Graetz, C. Lee, E. Muelly, C. Kennedy, E. Kim, Patient characteristics associated with choosing a telemedicine visit vs office visit with the same primary care clinicians. JAMA Netw. Open **3**(6), e205873 (2020). https://doi.org/10.1001/jamanetworkopen.2020.5873
136. Report of the WCPT/INPTRA Digital Physical Therapy Practice Task Force (2019).
137. A. Riazi, A.J. Thompson, J.C. Hobart, Self-efficacy predicts self-reported health status in multiple sclerosis. Mult. Scler. (Houndmills, Basingstoke, England) **10**(1), 61–66 (2004)
138. K.P. Rimmer, M. Kaminska, M. Nonoyama, E. Giannouli, F. Maltais, D.L. Morrison, et al., Home mechanical ventilation for patients with Amyotrophic Lateral Sclerosis: A Canadian Thoracic Society clinical practice guideline. Can. J. Respir. Crit. Care Sleep Med. **3**(1), 9–27 (2019). https://doi.org/10.1080/24745332.2018.1559644
139. M.O. Roland, R.W. Morris, A study of the natural history of back pain. Part 1: Development of a reliable and sensitive measure of disability in low back pain. Spine **8**, 141–144 (1983)
140. S. Rooney, D.A. McFadyen, D.L. Wood, D.F. Moffat, P.L. Paul, Minimally important difference of the fatigue severity scale and modified fatigue impact scale in people with multiple sclerosis. Mult. Scler. Relat. Disord. **35**, 158–163 (2019). https://doi.org/10.1016/j.msard.2019.07.028
141. J.M. Rothstein, Impairments: Always linked to meaningful disability? Phys. Ther. **81**, 886–887 (2001)
142. P. Sandstedt, S. Johansson, C. Ytterberg, C. Ingre, L.W. Holmqvist, M. Kierkegaard, Predictors of health-related quality of life in people with amyotrophic lateral sclerosis. J. Neurol. Sci. **370**, 269–273 (2016). https://doi.org/10.1016/j.jns.2016.09.034

143. R. Sayeed, D. Gottlieb, K.D. Mandl, SMART markers: Collecting patient-generated health data as a standardized property of health information technology. NPJ Digit. Med. **3**(1), 1–8 (2020). https://doi.org/10.1038/s41746-020-0218-6
144. K. Scalise, D.D. Allen, Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform. Br. J. Math. Stat. Psychol. **68**, 478–496 (2015). https://doi.org/10.1111/bmsp.12057
145. J.C. Seida, C. LeBlanc, J.R. Schouten, S.S. Mousavi, L. Hartling, B. Vandermeer, … D.M. Sheps. (2010). Systematic review: Nonoperative and operative treatments for rotator cuff tears. Ann. Int. Med. https://doi.org/10.7326/0003-4819-153-4-201008170-00263
146. A. Shumway-Cook, M. Woollacott, *Motor Control: Translating Research into Clinical Practice*, 4th edn. (Lippincott Williams & Wilkins, Philadelphia, 2012)
147. Social Security Administration. *Disability Benefits* (2019). Retrieved January 09, 2021, from https://www.ssa.gov/pubs/EN-05-10029.pdf
148. E. Spiegel, S. Jondhale, I. Brajkovic, K.C. Nesbit, I.E. Allen, V. Bhutani, et al., Valuation of quality of life in pediatric disability in a developing country. J. Child Neurol. **33**(9), 601–609 (2018). https://doi.org/10.1177/0883073818773941
149. N.B. Stergiou, M.J. Kurz, J. Heidel, Nonlinear tools in human movement, in *Innovative Analysis for Human Movement*, ed. by N. B. Stergiou, (Human Kinetics, Champaign, 2004), pp. 63–90
150. N. Stergiou, R.T. Harbourne, J.T. Cavanaugh, Optimal movement variability: A new theoretical perspective for neurologic physical therapy. J. Neurol. Phys. Ther. **30**, 120–129 (2006)
151. M. Sterling, D. Brentnall, Patient specific functional scale. Aust. J. Physiother. **53**(1), 65 (2007)
152. P. Stolee, K. Stadnyk, A.M. Myers, K.J. Rockwood, An individualized approach to outcome measurement in geriatric rehabilitation. J. Gerontol. Ser. A Biol. Sci. Med. Sci. **12**, M641–M647 (1999)
153. N.L. Stout, How registry data will change our approach to lymphedema research and clinical management. Rehabil. Oncol. **36**(1), 73–75 (2018). https://doi.org/10.1097/01.REO.0000000000000108
154. P. Stratford, C. Gill, M. Westaway, J. Binkley, Assessing disability and change on individual patients: A report of a patient specific measure. Physiother. Canada **47**, 258–263 (1995)
155. K. Suh, J. Beck, W. Katzman, D.D. Allen. Homelessness and rates of physical dysfunctions characteristic of premature geriatric syndromes: Systematic review and meta-analysis. Physiother. Theory Pract. (2020). https://doi.org/10.1080/09593985.2020.1809045
156. E.J. Sullivan. *Clinical Trial Endpoints* (2010). Retrieved from fda.gov/media/84987/download
157. N. Tellez, J. Rio, M. Tintore, C. Nos, I. Galan, X. Montalban, Does the modified fatigue impact scale offer a more comprehensive assessment of fatigue in MS? Mult. Scler. **11**(2), 198–202 (2005)
158. Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease. (2012). In Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. https://doi.org/10.17226/13284
159. U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, & Center for Devices and Radiological Health. (2009). Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Suppor Labeling Claims. Retrieved from fda.gov/media/77832/download
160. L.M. Verbrugge, A.M. Jette, The disablement process. Soc. Sci. Med. **38**, 1–14 (1994)
161. D.A. Vyas, L.G. Eisenstein, D.S. Jones, Hidden in plain sight–reconsidering the use of race correction in clinical algorithms. New Engl. J. Med. **383**, 874–882 (2020). https://doi.org/10.1056/NEJMms2004740
162. T.D. Wade, Traits and types of health data repositories. Health Inf. Sci. Syst. **2**(1), 1–8 (2014)

163. J.E. Ware, B. Gandek, Overview of the SF-36 health survey and the international quality of life assessment (IQOLA) project. J. Clin. Epidemiol. **51**(11), 903–912 (1998)
164. G. Wells, D. Beaton, B. Shea, M. Boers, L. Simon, V. Strand, et al., Minimal clinically important differences: Review of methods. J. Rheumatol. **28**(2), 406–412 (2001)
165. M. Wilson, *Constructing Measures: An Item Response Modeling Approach* (Erlbaum, Mahwah, 2005)
166. M. Wilson, D.D. Allen, J.C. Li, Improving measurement in behavioral sciences using item response modeling: Introducing item response modeling. Health Educ. Res. **21**(Supplement 1), i4–i18 (2006)
167. A. Wright, J. Hannon, E.J. Hegedus, A.E. Kavchak, Clinimetrics corner: A closer look at the minimal clinically important difference (MCID). J. Man. Manip. Ther. **20**, 160–166 (2012)
168. J. Yeskoo, J. Lee, C. Hoang, D.D. Allen, Effectiveness of risk-stratified care for low back pain using the subgroups for targeted treatment back screening tool: An evidence-based review and meta-analysis [abstract]. J. Orthop. Sports Phys. Ther. **48**(1), A26–A27 (2018)

Chapter 5
An Adaptive Strategy for Measuring Patient-Reported Outcomes: Incorporating Patient Preferences Relevant to Cost-Benefit Assessments of Vision Rehabilitation

Robert W. Massof and Chris Bradley

Abstract Vision rehabilitation aims to improve daily functioning in patients with chronic disabling vision impairments – low vision – by providing vision assistive equipment and teaching patients effective use of the equipment; teaching patients adaptations to obviate dependence on vision; modifications of the patient's environment to improve visibility and increase safety; and psychosocial counseling and education to help patients cope with the stress of impaired vision. The interventions that constitute vision rehabilitation are tailored to each patient's functional goals and personal preferences. This chapter describes the theory, structure, and mechanics of an adaptive patient-reported outcome measure – the Activity Inventory (AI). The AI was developed to identify each patient's unique set of functional *goals* within a hierarchical framework, to specify the cognitive and motor activities (*i.e.*, *tasks* in different functional domains) the patient customarily performs to achieve his or her personal goals, and to elicit ordinal patient ratings of the importance and difficulty of each goal and the difficulty of relevant subsidiary tasks under goals that meet or exceed defined importance and difficulty criteria. The theoretical assumptions of a Rasch psychometric model that is employed to estimate measures of the patient's overall visual ability from goal difficulty ratings and ability in each functional domain from task difficulty ratings before and after vision rehabilitation are presented and the results for a large group of low vision patients are reviewed. This chapter then explores the issue of how individual patient preferences must be factored into the outcome measures to determine the utility of functional improvements from vision rehabilitation. A theoretical framework for incorporating the utility of individual vision rehabilitation outcomes is offered and its application to functional outcome data is demonstrated.

R. W. Massof (✉) · C. Bradley
Wilmer Eye Institute, Johns Hopkins University, Baltimore, MD, USA
e-mail: bmassof@jhmi.edu

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3_5

Keywords Low vision rehabilitation · Personalized care · Adaptive instrument administration

5.1 Introduction

Rasch rating scale models have three free parameters to be estimated from rating scale questionnaire responses: (1) person measures; (2) item measures; and (3) rating category thresholds. When developing and validating patient-reported outcome measures (PROM), practitioners of Rasch analysis typically regard differential item functioning (DIF), differential person functioning (DPF), mistargeting of item measures to person measures, and disordering of rating category thresholds to be distressing. These indicators, taken at face value, are interpreted as signs of metrological trouble with the data that warrant modifying the list of items and rating categories to improve measurement purity.

In this chapter, we describe our approach to developing a PROM, the *Activity Inventory* (AI), for adaptively measuring the outcomes of rehabilitation of daily functioning by people with chronic disabling vision impairments – low vision. Vision rehabilitation aims to help people with chronic vision impairments overcome vision disabilities through behavioral and environmental modifications, the use of vision assistive equipment, education, and psychosocial counseling. The multifaceted and idiosyncratic nature of vision rehabilitation poses significant challenges to measuring its outcomes. Consequently, with respect to the tenets of Rasch analysis (specific objectivity, sufficiency of raw scores, and invariant comparison), our approach to overcoming the challenges to measuring outcomes of vision rehabilitation may seem iconoclastic. The most heretical features of our approach are:

- Vision rehabilitation requires different interventions designed to target different activities. Because patients respond to different subsets of items and the different daily activities define the content of different items in the AI, a positive outcome of vision rehabilitation is expected to manifest as *intervention-specific DIF*, which when taken literally appears to challenge specific objectivity.
- Official definitions of vision disability traditionally have been based on a criterion visual acuity in the better seeing eye [1] (e.g., ≤20/200 [USA] or <3/60 [WHO]) that cannot be improved with eyeglasses or contact lenses) or by peripheral vision loss specified by the horizontal extent of the visual field (e.g., ≤20° [USA] or ≤10° [WHO]). Given that these two types of vision impairment can occur separately, or in varying degrees together, and they differentially affect the person's ability to perform different activities that define the AI item content, we expect to see *DPF*, which also appears to challenge specific objectivity.
- The endpoint of vision rehabilitation is the attainment of activity-performance goals that are defined and prioritized by the individual's personal preferences. Consequently, the AI items must be administered adaptively, drawn from a calibrated item bank having anchored item measures. However, the choice of

items is driven by the visually impaired person's preferences, not by item administration efficiency to achieve a criterion level of precision in the estimation of the person measure. Thus, the items drawn from the item bank parallel the content of the *individualized rehabilitation plan* and, unlike the strategy of computer adaptive testing (CAT), are *not necessarily well-targeted to the neighborhood of the individual's person measure.* Contrary to this strategy, students of Rasch analysis are taught that a well-designed instrument has a rectangular distribution of item measures that spans the distribution of person measures, or at least an item measure distribution that matches the person measure distribution so that measurement precision is highest where the density of persons to be discriminated is highest [2].

- The respondent to the AI is asked to rate the difficulty of performing activities described by the item content. Particularly in light of intervention-specific DIF, to be measurable the outcome of vision rehabilitation for each activity must manifest as a change in the difficulty rating of the corresponding item. The responsiveness of the AI to a change in difficulty depends on the number of rating categories and on the sizes of the concatenated rating category intervals, which are separated by category thresholds. However, the polytomous Rasch models most often used to estimate measures from responses to rating scale questionnaires routinely estimate disordered thresholds, especially as the number of rating categories is increased in an attempt to improve resolution on the measurement scale. Because *disordered rating category thresholds are illogical and an unacceptable output of a valid rating scale instrument,* it has been assumed that the problem is with the data. Consequently, instrument developers are advised to reduce the number of rating categories after the fact through mergers of neighboring categories until the offending threshold disordering is eliminated [2]. A more rational approach is to assume that respondents understand what "ordered" means and the analyst employs a *measurement model in which proper ordering of category thresholds is axiomatic.*

5.2 A Person-Centered Measure for Vision Rehabilitation

The primary aim of vision rehabilitation is to improve the visually impaired person's functional ability on an activity-by-activity basis by ameliorating functional limitations caused or exacerbated by the person's vision impairment. Functional ability is a multidimensional construct (*e.g.*, cognitive, motor, psychological, sensory, etc.). Vision rehabilitation targets one dimension of functional ability – visual ability (*i.e.*, the ability to perform activities that depend on vision), which in turn may be multidimensional given the different types of vision impairments inherent in official definitions of vision disability that can occur independently (*e.g.,* reduced visual acuity, visual field loss, impaired color discrimination).

5.2.1 A Measurement Model for Visual Ability

For visually impaired people, each activity described by items in a visual function rating scale questionnaire [3] requires some amount of visual ability to be performed – a fixed latent variable called the *item measure*. In this chapter, we will denote latent variables with Greek letters and manifest variables (as well as constants, indices, functions, and operators) with Latin letters. Accordingly, we use the lower-case Greek letter rho and the subscript j to represent the average amount of visual ability required by the visually impaired population to perform the jth item (ρ_j). However, the actual amount of visual ability required by an item can vary randomly between persons depending on available technology, customary practices, environmental factors, etc. Thus, for the nth person, the actual visual ability required to perform the jth item ($\rho_{n,\,j}$) is expected to deviate from the average item measure for the population of interest by an amount for each person that is randomly distributed in the specified population ($\varepsilon_{n,\,j}$), such that $\rho_{n,\,j} = \rho_j + \varepsilon_{n,\,j}$. By definition, $\varepsilon_{n,\,j}$ can be positive or negative and the mean of epsilon values across all persons is zero for each item. Ignoring within-person variance in the deviate for the time being, the standard deviation of epsilon between persons for the jth item is denoted as σ_j (although technically the standard deviation is not a latent variable, it is referring to the distribution of a latent variable, and sigma is the conventional notation for standard deviation).

Each person has some amount of visual ability – a trait of the person represented by a latent variable called the *person measure*. Thus, α_n denotes the amount of visual ability possessed by the nth person. If person n has far more ability (the value of alpha) than is required to perform the activity described by item j, *i.e.*, $\alpha_n > \rho_{n,\,j}$, then the person is likely to report that the activity is "not difficult" to do. If person n has far less ability than is required to perform the activity described by item j, *i.e.*, $\alpha_n < \rho_{n,\,j}$, then the person is likely to report that the activity is "impossible" to do. Between these two extremes, person n could use an ordinal rating scale to estimate the level of difficulty she or he has performing the activity described by item j (*e.g.*, 1 – "very difficult"; 2 – "moderately difficult"; 3 – "somewhat difficult"; and to complete the scale we would add the extremes: 4 – "not difficult" and 0 – "impossible to do"). Theoretically, we interpret the chosen rating as the person's magnitude estimate of his or her *functional reserve* for the activity described by the item content [4]. Functional reserve, denoted as $\varphi_{n,\,j}$ for person n relative to item j, is simply the difference between the person measure and the measure for the item that person is responding to, $\varphi_{n,\,j} = \alpha_n - \rho_{n,\,j}$. From the definition of $\rho_{n,\,j}$, functional reserve also can be written as $\varphi_{n,\,j} = \alpha_n - \rho_j - \varepsilon_{n,\,j}$, which incorporates the average item measure for the specified population, ρ_j, and the deviation from the average for person n, $\varepsilon_{n,\,j}$.

We can now think of creating a functional reserve ruler in units of the latent variable phi. Although all persons are given the same ordinal difficulty rating categories, each person divides the φ ruler into his or her own set of intervals – the only thing persons agree on is that the interval for rating category 0 comes before and is concatenated with the interval for rating category 1, which comes before and is concatenated with the interval for rating category 2, etc. Although the sizes and locations of these intervals on the φ ruler are likely to be unique to the person, they must be in the order of the φ magnitude estimates they represent and separated from their neighboring intervals by boundaries located at positions unique to the person. The boundaries between ordered intervals on any given trial are called response *thresholds*, and their locations in φ units for the example we have been using are $\tau_{n,\ 1}$, $\tau_{n,\ 2}$, $\tau_{n,\ 3}$, $\tau_{n,\ 4}$ for person n. The φ scale is open-ended so that the lower bound for interval 0 is negative infinity and the upper bound for interval 4 is positive infinity. People most likely do not agree with one another, and each person may be inconsistent from trial to trial in the positions of the different response thresholds on the phi ruler, but by definition the thresholds must be ordered on every trial.

As we did with the item measures, we can estimate a population-based average threshold for each interval, which are fixed values, and define the threshold for the xth interval as the average threshold (τ_x) plus a person-specific deviate ($\eta_{n,\ x}$) that is randomly distributed between persons, *i.e.*, $\tau_{n,\ x} = \tau_x + \eta_{n,\ x}$, for $x = 1$ to 4. The average value of eta across persons for each threshold is zero and the standard deviation of eta between persons is σ_x. With these definitions and explanations, the use of a rating scale to make magnitude estimates implies that for person n to assign a rating of x to item j, functional reserve ($\varphi_{n,\ j}$) must be greater than or equal to person n's threshold for category x and less than person n's threshold for category $x + 1$, and all intervals must be ordered:

$$\cdots < \tau_{n,x-1} < \tau_{n,x} \le \varphi_{n,j} < \tau_{n,x+1} < \tau_{n,x+2} < \cdots$$

Stated more precisely, we assume a real line is partitioned by ordered thresholds into ordered intervals called rating categories, where the rating categories are defined using half-open intervals so that every point on the real line corresponds to precisely one rating category. Substituting terms in this expression with sums of fixed variables and randomly distributed deviates, we obtain

$$\cdots < \tau_{x-1} + \eta_{n,x-1} < \tau_x + \eta_{n,x} \le \alpha_n - \rho_j - \varepsilon_{n,j} < \tau_{x+1} + \eta_{n,x+1} < \tau_{x+2} + \eta_{n,x+2} < \cdots$$

The deviate $\varepsilon_{n,\ j}$ can be added to each term in the above expression, and defining a new deviate, $\zeta_{n,\ j,\ x} = \eta_{n,\ x} + \varepsilon_{n,\ j}$, that is randomly distributed between persons, we obtain

$$\cdots < \tau_{x-1} + \zeta_{n,j,x-1} < \tau_x + \zeta_{n,j,x} \le \alpha_n - \rho_j < \tau_{x+1} + \zeta_{n,j,x+1} < \tau_{x+2} + \zeta_{n,j,x+2} < \cdots \tag{5.1}$$

But there are two ways to interpret the error term, $\zeta_{n,\ j,\ x}$, on the average rating category thresholds used in expression (5.1): (1) the threshold error term can be identified as a deviate or (2) it can be identified as a random variable. As a deviate we can describe the trial-to-trial distribution of threshold values while enforcing threshold ordering [5], whereas random variables can take on any value, including those that result in threshold disordering. To represent this second interpretation of rating category thresholds, expression (5.1) is modified to be less specific about the ordering of thresholds:

$$\left\{\cdots,\tau_{x-1}+\zeta_{n,j,x-1},\tau_x+\zeta_{n,j,x}\right\}\leq\alpha_n-\rho_j<\left\{\tau_{x+1}+\zeta_{n,j,x+1},\tau_{x+2}+\zeta_{n,j,x+2},\cdots\right\} \tag{5.2}$$

Expression (5.2) says that for person n to respond with rating category x on a given trial, functional reserve must be greater than or equal to all the average thresholds (plus a random variable) on the left and be less than all of the average thresholds (plus a random variable) on the right (*i.e.*, on each trial the person's thresholds are segregated into two subsets, one in which all elements do not exceed the functional reserve and the other in which all elements are greater than the functional reserve). Expression (5.1) is simply one form that the more general expression (5.2) can take.

Mathematically, the assumptions built into expression (5.1) require average thresholds to be ordered whereas the assumptions built into expression (5.2) permit average thresholds to be disordered. In other words, expression (5.1) emphasizes the intervals – they must be ordered and concatenated on every trial for every person, but their sizes and how they are centered as a group on the φ scale are free to vary between persons and between trials by the deviate zeta. For expression (5.1), the locations of the response category thresholds are identified as the boundaries of the concatenated intervals. Thus, $\tau_{x-1}+\zeta_{n,j,x-1}$ is always less than $\tau_x+\zeta_{n,j,x}$, which in turn is always less than $\tau_{x+1}+\zeta_{n,j,x+1}$, etc.

For expression (5.2), the definition of "threshold" is subtly changed, even though when casually described it continues to be used as if it refers to the boundaries of an interval in which φ_{nj} falls. But for expression (5.2) the ordinal value assigned to the response refers to a count of the number of thresholds pre-labeled with ordered numbers that are less than $\varphi_{n,j}$. That is, on each trial for each person the response category thresholds are identified and permanently labeled (like a number on a soccer player's jersey). The numbers on the thresholds represent the order in which they are counted, not necessarily the order of the threshold magnitudes.

Mathematically, to satisfy expression (5.2) it is not necessary for thresholds to define boundaries of intervals – each threshold is tracked from trial to trial on the φ scale as an independent entity. When the person responds with rating x, it is assumed by the model that $\varphi_{n,j}$ is greater than all thresholds less than $\tau_{x+1}+\zeta_{n,j,x+1}$ and less than all thresholds greater than $\tau_x+\zeta_{n,j,x}$. Trials that do not satisfy expression (5.2) are ignored (*i.e.*, estimated probabilities are conditioned on the requirement that dichotomous scores assigned to each threshold satisfy a Guttman scale on each trial) [6]. Unlike the requirement for expression (5.1), on each eligible trial the two sets of thresholds plus error segregated by $\varphi_{n,j}$ can have magnitudes in any order on their respective sides of the inequalities. For this reason, Rasch models derived from expression (5.2) can estimate disordered average thresholds, whereas Rasch models derived from expression (5.1) always estimate ordered thresholds [5, 7, 8].

Some Rasch models permit expected values of thresholds for response categories to vary across items (*i.e.*, τ_x is redefined as $\tau_{j,\ x}$) [9]. If expression (5.1) must be satisfied by $\tau_{j,\ x}$ (*i.e.*, thresholds for each item must be ordered), expression (5.1) can be rewritten as

$$\cdots < \tau_{j,x-1} + \rho_j + \zeta_{n,j,x-1} < \tau_{j,x} + \rho_j + \zeta_{n,j,x} \leq \alpha_n < \tau_{j,x+1} + \rho_j + \zeta_{n,j,x+1} < \tau_{j,x+2} + \rho_j + \zeta_{n,j,x+2} < \cdots \tag{5.3}$$

by adding ρ_j to every threshold and to functional reserve, which in effect, creates a different Likert scale for each item (j). Expression (5.3) is consistent with Samejima's [10] graded response model if the variance of $\zeta_{n,\ j,\ x}$ depends on the item (j) – a common assumption of item response theory models. If the variance of $\zeta_{n,\ j,\ x}$ is constant with respect to n, j, and x, as required by Rasch models [8], then expression (5.3) reduces to a dichotomous Rasch model with $\tau_{j,\ x} + \rho_j$ defining the measure for "item" x in domain j [6]. Using rating scale questionnaire terminology, the $\tau_{j,\ x}$ values are playing the role of $x = 1$ to m_j dichotomously scored items (j,x) and the ρ_j values are playing the role of domain-dependent (j) constants that offset the item measures in domain j so that all items in the instrument share the same origin on a common scale. The scoring rule for expressions (5.1, 5.2 and 5.3) is the same, which can be stated explicitly as: if the respondent chooses category j,x, then response category j,x and all categories less than j,x are given a score of 1 and all categories greater than j,x are given a score of 0 [6, 11, 12]. This rule is the consequence of requiring normative measurement models to conform to a Guttman scale.

Unlike Rasch's original application of his model to reading tests [12] and most current applications in educational testing, in which ordinal scores represent error counts, ordinal scores for rating scales are assigned to subjective magnitude estimates, which are represented by a distance between two points on a number line [13]. The only counts involved in a distance measure are counts of unit distances that are concatenated to span the distance being measured, not counts of events. Expression (5.1) (and algebraic variations) represents a class of Rasch models for which measurement units correspond to a distance on a number line. Expression (5.2) (and algebraic variations) represents a class of Rasch models for which measurement units correspond to counts of sequentially ordered events that are positioned independently at points on the number line [6].

If $r_{j,\ x}$ is the correlation between $\eta_{n,\ x}$ and $\varepsilon_{n,\ j}$, then the between-person variance in $\zeta_{n,\ j,\ x}$ is $\sigma^2_{j,x} = \sigma^2_j + \sigma^2_x + 2r_{j,x}\sigma_j\sigma_x$. Invoking the central limit theorem, we assume the probability density function for $\zeta_{n,\ j,\ x}$ is approximately normal, which in turn is further approximated with the logistic density function: $f\left(\zeta_{n,j,x}|0,\sigma_{j,x}\right) = \frac{e^{-\zeta_{n,j,x}/\sigma_{j,x}}}{\sigma_{j,x}\left(1+e^{-\zeta_{n,j,x}/\sigma_{j,x}}\right)^2}$.

From expression (5.1), the probability person n will rate item j with rating category

x is the probability $\tau_x + \zeta_{n,j,x} < \alpha_n - \rho_j$ and $\alpha_n - \rho_j < \tau_{x+1} + \zeta_{n,j,x+1}$. To satisfy this requirement in expression (5.1), it must be true that $\tau_{x+1} + \zeta_{n,j,x+1} > \tau_x + \zeta_{n,j,x}$ for every person/item combination and for every value of x.

For dichotomous response categories, there is only one threshold, which can be added to ρ_j and its variance can be added to σ_j^2. The major difference between dichotomous Rasch models and most dichotomous item response theory (IRT) models is that dichotomous Rasch models assume $\sigma_j = 1$ for all items, whereas IRT models typically estimate σ_j for each item. This difference means that Rasch models are normative measurement models (*i.e.*, modeling the measurement, not the data) and IRT models are descriptive statistical models (modeling the data, not the measurement). This difference carries through to polytomous rating scale models also – Rasch models assume $\sigma_{j,x} = 1$ for all items and thresholds and most IRT models assume $\sigma_{j,x} = \sigma_j$ for all items and thresholds – with the same consequences for how the models are characterized. Our aim is to estimate *measures* of visual ability from item difficulty ratings, so we employ a Rasch model, which means that the probability density function for measurement errors is $f(\zeta_{n,j,x}|0, \sigma_{j,x}) = f(\zeta|0, 1)$. A result of this assumption is that all rating scale responses estimated by the model must adhere to a Guttman scale, a *sine qua non* of measurement. In effect, Rasch models estimate person and item measures that generate the most likely Guttman scale underlying the observations, whereas IRT models estimate parameters that generate responses that best fit the data [8].

Again returning to expression (5.1) with the assumption that $\sigma_{j,x} = 1$, we see that the probability person n responds to item j with ***category x* or greater** is $p(\tau_x + \zeta_{n,j,x} < \alpha_n - \rho_j)$, which is equal to $p\left(\zeta_{n,j,x} < \alpha_n - \rho_j - \tau_x\right) = \int_{-\infty}^{\alpha_n - \rho_j - \tau_x} f(\zeta|0,1)d\zeta$, a dichotomous Rasch model. Similarly, we see from expression (5.1) that the probability person n responds to item j with ***less than category x + 1*** is $p(\alpha_n - \rho_j < \tau_{x+1} + \zeta_{n,j,x+1})$, which also is in the form of a dichotomous Rasch model, $p\left(\zeta_{n,j,x+1} > \alpha_n - \rho_j - \tau_{x+1}\right) = \int_{\alpha_n - \rho_j - \tau_{x+1}}^{\infty} f(\zeta|0,1)d\zeta = 1 - \int_{-\infty}^{\alpha_n - \rho_j - \tau_{x+1}} f(\zeta|0,1)d\zeta$.

The probability person n responds with category x to item j is

$$p\left(x|\alpha_n - \rho_j\right) = 1 - \left(p\left(\zeta_{n,j,x} < \alpha_n - \rho_j - \tau_x\right) + p\left(\zeta_{n,j,x+1} > \alpha_n - \rho_j - \tau_{x+1}\right)\right) \quad (5.4)$$

which is in the Rasch model form ($\sigma_{j,x} = 1$) of Muraki's modification of Samejima's graded response model [10, 14]. The three polytomous Rasch model parameters **—** $\alpha_n$for each person, ρ_j for each item, and τ_x for each threshold – are estimated for the logistic difference model in Eq. (5.4) [15] using the *method of successive dichotomizations* [5].

5.2.2 *Defining and Organizing the Activity Inventory Item Content*

Because of the demographics of low vision [16], vision rehabilitation primarily targets older visually impaired adults [17]. The low vision patient evaluation typically begins with an intake interview that consists of a health history, visual impairment history, psychosocial history, and functional history [18]. The functional history itemizes how the patient's visual impairment limits her/his ability to live independently, ability to engage in social activities, and ability to engage in favored leisure activities and avocations. For younger patients, the functional history might also cover limitations on employment-related and/or school-related activities, however because most low vision is due to age-related eye diseases, those patients are rare and tend to be referred out of the health care system for vocational rehabilitation or special education services.

The Activity Inventory (AI) was developed both to structure the intake history, so as to facilitate the development of individualized rehabilitation plans, and to provide an adaptive visual function rating scale questionnaire for measuring the low vision patient's functional ability and outcomes of vision rehabilitation [19, 20]. A retrospective chart review of low vision patient intake histories identified 460 common cognitive and motor activities that frequently were mentioned by patients as important for them to be able to perform but were made unusually difficult or precluded by their visual impairments. These very specific activities, called "Tasks" in the AI, were grouped according to the activity "Goals" they serve [19]. Goals in the AI refer to the reason for performing a coordinated set of Tasks (the first letter in the terms "Goals" and "Tasks" is capitalized when referring to items in the AI). For example, "prepare daily meals" is a Goal. There are many ways to prepare daily meals ranging from performing a suite of customary Tasks required to prepare a meal from scratch (*e.g.*, read recipes, measure ingredients, cut food, adjust appliance controls, time cooking, judge doneness of food, etc.) to heating up prepared food in a microwave or conventional oven (*e.g.*, read package instructions and adjust appliance controls).

Two rehabilitation strategies are employed to achieve activity Goals: (1) make usual and customary Tasks less difficult to perform by using vision assistive equipment to enhance vision (*e.g.,* magnifier) or using sensory substitution technology to obviate vision (*e.g.,* bump dots on appliance controls, talking timer, electronic liquid level indicator), or (2) employ adaptive strategies to make it possible to achieve the Goal without having to perform the patient's usual and customary Tasks (*e.g.*, heat prepared food instead of cooking from scratch). Overall outcomes of vision rehabilitation generally are judged in terms of goal attainment, whereas the effectiveness of vision assistive equipment and visual skills instruction tends to be judged in terms of improvements in the performance of tasks.

The difference between education and rehabilitation can be captured by the difference between learning to cook and regaining the ability to cook. Rehabilitation goals are defined by lost functional ability and by patient preferences. However, how much value the patient places on an activity depends on the Objective of the Goal.

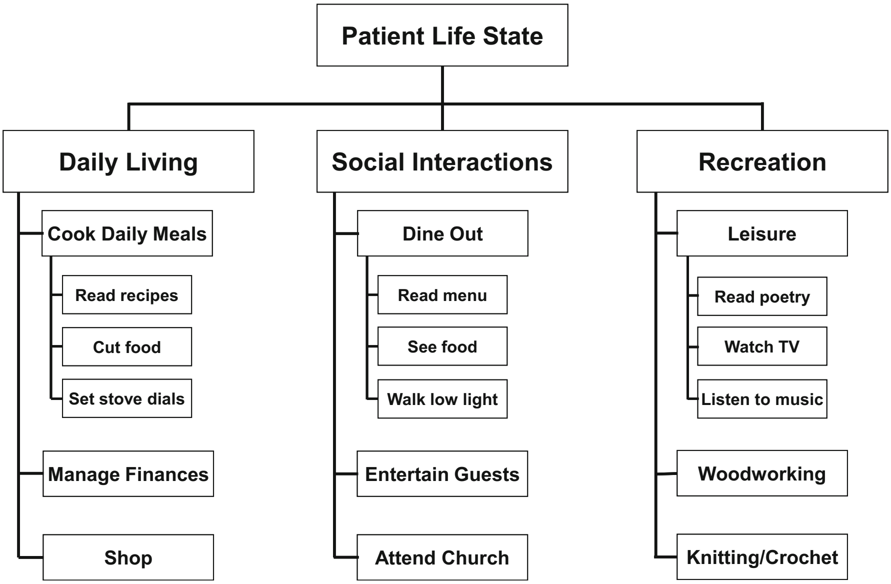


Fig. 5.1 Schematic of the hierarchical Activity Breakdown Structure (ABS) for the AI. Activities can be grouped into three Objectives that define the patient's life state: Daily Living, Social Interactions, and Recreation (vocation and education are not included as objectives that are addressed by vision rehabilitation services provided within the health care system). Under each Objective are specific activity Goals (*e.g.*, Cook Daily Meals under Daily Living) that often are identified by low vision patients as needed to meet the parent Objective (the AI has 50 Goals: 18 under Daily Living, 11 under Social Interactions, and 21 under Recreation – Goals rated "not important" or "not difficult" are not included in the ABS). The AI has a total 460 Tasks in its calibrated item bank that are nested under the Goals they serve. Tasks rated by the patient as "not relevant" or "not difficult" are not included in the ABS. Besides being nested under Goals, each Task is also assigned to one of 4 functional domains: Reading; Mobility; Visual Motor (i.e., eye-hand coordination); or Visual Information (*i.e.*, perception)

For example, cooking to prepare daily meals might be an extremely important Goal to the patient because it serves the Objective of daily living (*i.e.,* necessary for the patient to live independently). Cooking meals also could serve a social interaction Objective that is important to the patient because the patient places high value on entertaining guests. And cooking could serve a recreation Objective because the patient is rewarded by the joy of cooking. Although heating prepared food in a microwave oven might be an acceptable adaptation if cooking is serving the Daily Living Objective, it might not be acceptable if it is serving the Social Interaction or Recreation Objectives. As schematized by the example in Fig. 5.1, there are 50 common Goals in the AI that are nested under the three Objectives. Borrowing terminology from project management, this hierarchical structure of AI Objectives, Goals, and Tasks is called the *Activity Breakdown Structure* (ABS) [19]. The ABS organizes the items in the AI (Goals and Tasks) in a way that parallels the functional history.

The AI Tasks also are organized by grouping them into four sets according to the type of rehabilitation intervention that might be used. Each set of Tasks is identified as a functional domain: (1) reading function, (2) mobility function, (3) visual motor function, and (4) visual information processing function. Interventions for reduced reading function include various forms of magnification, speech output optical character recognition apps, synthetic speech devices, audio recordings, and braille. Interventions for reduced mobility function include orientation and mobility training (*e.g.*, white cane use), dog guide, navigation apps (*e.g.,* GPS systems), ride share services, talking signs, and remote sighted guide services. Interventions for reduced visual motor function include signature guides, raised line guide for writing, needle threader, syringe fillers, vegetable cutters, oversize keyboards, bump dots, oversize nail clippers with magnifier, and organizational skill instruction. Interventions for reduced visual information processing function include face and object recognition smartphone apps, descriptive audio for movies and events, color identification smartphone apps, and environmental modifications.

It is important to note, as illustrated above, most interventions that constitute vision rehabilitation consist of tools and methods that are specific to a narrow set of activities. The aim of each intervention is to reduce the difficulty of performing the troublesome activity (*e.g.*, liquid level indicators sound an alarm when liquid in a glass or cup reaches a criterion level). Theoretically, this piecemeal approach to vision rehabilitation translates to increasing functional reserve for each item by reducing the visual ability required by the item, *i.e.*, reducing ρ_j in the model [21]. Selective reductions in item measures because of piecemeal intervention translate to *intervention-specific DIF.*

5.2.3 Adaptive Administration of the AI

The present AI item bank consists of 50 Goals distributed under 3 Objectives (Daily Living, Social Interactions, Recreation) and 460 Tasks nested under the Goals. There is redundancy of item content within the list of 460 Tasks, however item measures may vary between different Tasks that have the same content depending on the Goal and Objective the Task is serving. Similarly, there is some redundancy in Goal content because Goals could serve more than one Objective. Returning to the cooking example, “prepare your daily meals” is a Goal under Daily Living, “prepare food for guests” is a Goal under Social Interactions, and “cook or bake for recreation” is a Goal under Recreation. Although the list of Tasks for each of these Goals is very similar with respect to content, they differ in their item measures due to varying performance criteria that must be met to satisfy the Objective served by the parent Goal.

Because the respondent is visually impaired, the AI is administered by interview, usually over the telephone with the assistance of a secure computer-assisted telephone interview (CATI) system. For the baseline interview (before vision rehabilitation services are rendered), the patient is asked to rate the importance of being able to attain without the assistance of another person one of the 50 Goals. The response choices are “not important”, “slightly important”, “moderately important”, or “very important”. If the patient chooses “not important”, the response is recorded and the interviewer moves on to the next Goal. If the patient chooses any of the other levels of importance, the response is recorded and the interviewer asks the patient how difficult it is to attain the Goal without the assistance of another person. The response choices are “not difficult”, “somewhat difficult”, moderately difficult”, “very difficult”, or “impossible to do”. The patient’s response is recorded. If the patient responded “not difficult”, then the interviewer advances to the next Goal. If the patient responded with any of the other difficulty rating categories, then the interviewer asks the patient to rate the difficulty of performing each of that Goal’s subsidiary Tasks using the same five difficulty rating categories as used with the Goal, or to respond the Task is “not applicable” to the respondent’s customary way of achieving the Goal.

After completing the rating of Tasks under the Goal, the interviewer moves to the next Goal and repeats all the same steps. This approach is considered adaptive because the patient’s preferences (importance ratings of Goals and applicability responses for Tasks) determine which items will have difficulty ratings elicited. Goals also must be rated at least “slightly difficult” to have difficulty ratings of its subsidiary tasks elicited. The rationale for this adaptive approach is that if a Goal is not important or not difficult at baseline, it will not be included in the individualized rehabilitation plan.

5.2.4 Properties of Estimated AI Item and Person Measures at Baseline

Generically, we refer to the estimated functional reserve, φ_{nj}, as *functional ability* of person n with respect to the content of item j. Functional ability is a multidimensional construct (*e.g.*, cognitive ability, motor ability, sensory ability, psychological ability, etc. are components of functional ability). Each functional ability dimension has its own multidimensional structure (*e.g.*, sensory ability can be expressed as visual ability, hearing ability, taste ability, etc.), and each component ability subdimension can be expressed further with its own dimensional structures (*e.g.*, visual ability can have independent components differentially affected by ophthalmic diseases: night vision, visual acuity, peripheral vision, color vision, etc.), all of which are properties of the person expressed at different levels of detail that must be inferred from observations by way of theory.

The person and item measures estimated from Rasch analysis of AI difficulty ratings correspond to the magnitude of an origin-bound functional ability vector in the multidimensional functional ability space. The direction of the vector is determined by item content and by traits of the people rating the items. For example, some vision-dependent items might describe activities, like reading medication instructions, that depend more heavily on cognitive ability than on motor ability and the reverse might be true for other items, such as signing a check. Also, some low vision patients might have cognitive impairments, whereas other low vision patients might have physical impairments that differentially influence their ratings of different items.

These differences combined with variations between items in the demand placed on vision and variations between people in vision impairment severity will give rise to variations in the direction and magnitude of the resultant vector. To the extent that variations in visual ability is the common denominator for all patients and items, we presume the magnitude of the average vector across items and persons represents *visual ability*. All else being equal, or at least randomly distributed, variations of vision impairment severity between persons and variations of demand on vision between items will give rise to variations in functional reserve, which can be characterized as variations in the magnitude of the operationally-defined "visual ability" vector in functional ability space. Variations between items in the direction of this visual ability vector depend on how much demand the items place on other functional ability dimensions (*e.g.*, cognitive demand, physical motor demand, psychological demand). Variations between persons in the direction of this visual ability vector depend on the types and magnitudes of other functional impairments the person may have (e.g., cognitive, physical motor, psychological disorders). Variations in measurements of "visual ability" depend on both the amount of deviation in vector direction and the magnitude of the deviated vector, which projects onto the defined visual ability vector, the magnitude of which corresponds to the estimated measures. We use this vector representation to guide our analyses of sources of variance and covariance in estimated person and item measures from AI difficulty ratings by people with low vision.

Item measures for the 510 Goals and Tasks in the AI were estimated using the method of successive dichotomizations [5] from the difficulty ratings of about 3600 low vision patients at pre-rehabilitation baseline [22]. Because activity Goals are attained by successfully performing some subset of their subsidiary Tasks, the difficulties of attaining Goals, and therefore their item measures, are expected to be a monotonically increasing function of the difficulty of performing their subsidiary Tasks. The left panel of Fig. 5.2 illustrates a scatterplot of mean Task item measures (on the ordinate) vs parent Goal item measures (on the abscissa), both specified as a difference from the mean of all Goal item measures and the mean of all average Task item measures respectively, along with the expected relationship given the respective means (red line) [23]. The Pearson correlation is 0.48. These results are consistent with the hypothesis that the difficulty of an activity Goal is inherited from the difficulties of the more specific activity Tasks that serve the Goal.

The Goal and Task item measures are estimated simultaneously on the same scale, therefore person measures estimated from Goal difficulty ratings should agree with person measures estimated from Task difficulty ratings. The right panel of Fig. 5.2 illustrates a scatterplot of person measures estimated from Task difficulty ratings (on the ordinate) versus person measures estimated from Goal difficulty ratings (on the abscissa) along with the expected identity relationship (red line) for approximately 3600 low vision patients at pre-rehabilitation baseline. The Pearson correlation is 0.71.

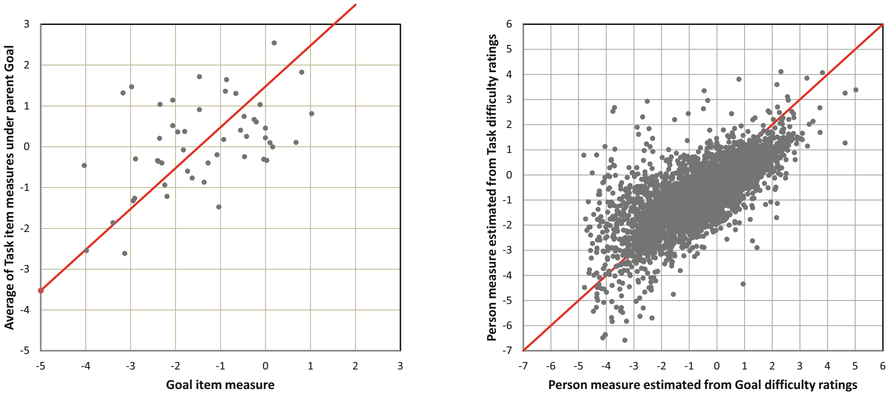


Fig. 5.2 ***Left panel***: Scatter plot of the average of all AI item measures of Tasks serving each Goal versus the corresponding Goal item measure. The red line is the identity line with respect to deviations of Goal measures from the mean Goal measure and deviation of the average Task measures from the mean of the average Task measures. Each data point corresponds to one of the 50 Goals. ***Right panel***: Scatter plot of person measures estimated from all Tasks rated by the person at baseline versus person measures estimated from all Goals rated by the same person at baseline. The red line is the identity line. Each data point corresponds to a separate person

Rasch models are normative measurement models. As opposed to most IRT models, which are descriptive statistical models of observations, Rasch models assume that the density function for random deviates, $f(\zeta|0, 1)$, is the same for every combination of persons and items. The validity of using a Rasch model to estimate measures from observations is tested by determining if the observations conform to the premises of the model (*i.e.,* determining if "the data fit the model" rather than fitting the model to the data). The *information-weighted mean square fit statistic* (infit) is used to test the validity of item and person measure estimates. Equation (5.4) is the probability that person n would respond with category x to item j. Thus, if $x_{n,\ j}$ is the observed response of person n to item j and $\mathbb{E}\left\{x_{n,j}|\alpha_n,\rho_j\right\} = \sum_{x=0}^{4} xp\left(x|\alpha_n - \rho_j\right)$ is the response of person n to item j expected by the model, then the sums of squares of observed minus expected responses across persons for each item is $SS_j = \sum_{n=1}^{N_j}\left(x_{n,j} - \mathbb{E}\left\{x_{n,j}|\alpha_n,\rho_j\right\}\right)^2$, where N_j is the number of persons who rated item j. The expected sums of squares for item j is $\mathbb{E}\left\{SS_j\right\} = \mathbb{E}\left\{x_{n,j}^2|\alpha_n,\rho_j\right\} - \mathbb{E}\left\{x_{n,j}|\alpha_n,\rho_j\right\}^2$, for which $\mathbb{E}\left\{x_{n,j}^2|\alpha_n,\rho_j\right\} = \sum_{x=0}^{4} x^2 p\left(x|\alpha_n - \rho_j\right)$. Assuming the source of variance is normally distributed, we expect SS_j to have a chi-square distribution with $N_j - 1$ degrees of freedom. The expected value of a chi-square distribution is its degrees of freedom (*df*). Thus, the ratio of SS_j to the expected value of SS_j, given the assumption of an underlying normal density function as the source of variance, should be $\frac{SS_j}{\mathbb{E}\left\{SS_j\right\}} = \frac{\chi^2}{df_j}$. This ratio is called the *infit*. If $df > 25$, a cube-root transformation of χ^2 is a good approximation to a normal distribution (*i.e.*, Wilson-Hilferty transform [24]).

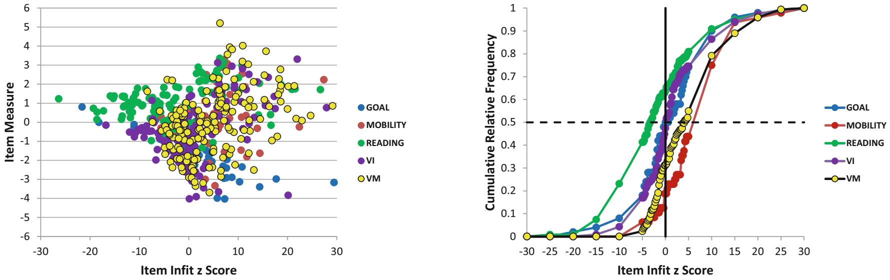


Fig. 5.3 ***Left panel***: Scatter plot of item measures for each of the 50 Goals and 460 Tasks as a function of the z-score for their respective item measure infit mean square. The Tasks are color coded by the functional domain to which they are assigned: Goals (blue), Mobility (red), Reading (green), Visual Information Processing (purple), and Visual Motor (yellow). The expected value for the infit z-score is 0. ***Right panel***: Cumulative frequencies of the infit z-scores for each functional domain as identified by color in the left panel

We estimated the infit for each of the 50 Goals and 460 Tasks in the AI from the responses of our large sample of low vision patients rating the difficulty of AI items and transformed the infits to standard normal deviates (z-scores) for each item. The results for all Goals and Tasks are illustrated in the left panel of Fig. 5.3 as the covariance of infit z-scores (abscissa) and item measures (ordinate) for each item. There is a very weak correlation between infit mean squares and item measures ($r = 0.12$). If the estimated measures conformed to the unidimensional assumption of the Rasch model (*i.e.*, magnitude of a single visual ability vector, $\varphi_{n,j}$) contaminated by a single source of normally distributed random error (ζ), 97% of the points would fall symmetrically about zero within a z-score range of ±2. Clearly, these assumptions are violated [25]. [N.B. Because the AI is adaptive, the number of persons who rated each item varied between items, so *df* must vary between items. The mean of a chi-square distribution is *df*, the variance is 2*df*, the skewness is $\sqrt{8/df}$, and the kurtosis is 12/*df*, thus variations in *df* across items result in variations in the shape of the composite infit distribution. However, because *df* is very large for all items, skewness and kurtosis are approximately zero. Although the Wilson-Hilferty transform can be used for each item infit, it is necessary to employ each item's respective df_j when transforming to z-scores.]

As described above, each Task in the AI is assigned to one of four functional domains: (1) reading, (2) mobility. (3) visual information processing, or (4) visual motor. The AI Goals and the functional domains to which the AI Tasks are assigned are color-coded in Fig. 5.3. It can be seen that the colors form different clusters of infit z-scores. The right panel of Fig. 5.3 shows cumulative frequencies of infit z-scores for Goals and for Tasks in each of the four functional domains. The infit z-score cumulative frequency functions are similar for each of the five groups of items, but the median z-score values (z-score corresponding to 0.5 cumulative frequency) vary between functional domains (the expected value of the median is

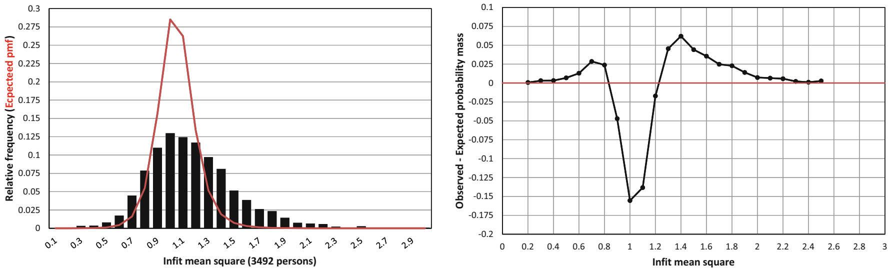


Fig. 5.4 ***Left panel***: Histogram (black bars) of the person measure infit mean square values. The red curve is the expected probability mass function (same bin width as the histogram) for the person measure infit mean square given the degrees of freedom (1 less than the number of items rated) for each patient. ***Right panel***: Difference between the observed (black bars in the left panel) and expected (red curve in the left panel) probability mass values for each infit mean square value. Differences greater than 0 indicate more variance observed in the distribution of person measure deviates than expected and differences less than 0 indicate less variance than expected

a z-score of 0 – solid vertical line). The median of the infit z-scores for Goals (0.2) and visual information processing Tasks (0.1) are very close to the expected value of zero. The median infit z-score for reading Tasks (−3.2) is much lower than the expected value of zero (indicative of error variance less than the expected amount of variance) and the median infit z-score for mobility Tasks (6) and visual motor Tasks (4.1) are much higher than the expected value of zero (indicative of error variance more than the expected amount of variance). These results suggest a strong dimensional structure to the estimated visual ability measure.

An infit also can be estimated for each person, but since the AI is administered adaptively, the number of items rated varies between people. The infit for person n is $\frac{SS_n}{\mathbb{E}\{SS_n\}} = \frac{\chi^2}{df_n}$, for which $df_n = J_n - 1$, where J_n is the number of items rated by person n. Thus, for adaptive testing, the expected infit frequency distribution is a sum of weighted chi-square distributions. The left panel of Fig. 5.4 illustrates a histogram of person infits (black bars). The red curve is the expected chi-square mixture probability mass function [26] (pmf – same bin width as the histogram) estimated for the 3600 respondents to the AI. The chi-square mixture pmf is the sum of weighted chi-square pmfs for different values of df_n with the weight equal to the fraction of persons who rated $df_n + 1$ items in the AI.

The right panel of Fig. 5.4 illustrates the difference between the observed and expected pmfs in the left panel of Fig. 5.4. About 27% of the persons had excess error variance in the observed responses (positive differences for infits greater than 1) and 8% had less error variance than expected (positive differences for infits less than 1). As discussed more formally in a later section, much of the excess variance may be due to functional limitations caused by comorbidities.

5.3 Functional Domains and Differential Person Functioning (DPF)

The analysis of item infit statistics by functional domains, as summarized in Fig. 5.3, suggests a strong dimensional structure to the estimated visual ability variable. Such a dimensional structure would correspond to DPF. Figure 5.5 shows that person measure distributions are significantly different when estimated from difficulty ratings of Goals and of Tasks in each of the four functional domains (ANOVA: $F = 30.95$, $df_B = 5$, $df_W = 20{,}502$, $p = 1.7 \times 10^{-31}$). Post hoc paired t-tests with Bonferroni adjustment for multiple comparisons showed that differences between all pairs are highly significant, except for visual information processing and visual motor functions ($p = 0.37$).

In a vector representation of covariances, vector magnitude corresponds to the square root of the variance (*i.e.*, standard deviation in units of the measure) explained by orthogonal factors and the cosine of the angle between any pair of vectors corresponds to the correlation between the variables those vectors represent. Such factors could represent visual components of the visual ability variable and/or contributions to estimated measures from other functional ability dimensions, such as cognitive disorders, physical motor limitations, and depressed psychological state.

Therefore, we employed factor analysis with principal axis factoring and varimax rotation (*i.e.*, optimizing the rotation of the orthogonal factors to maximize the

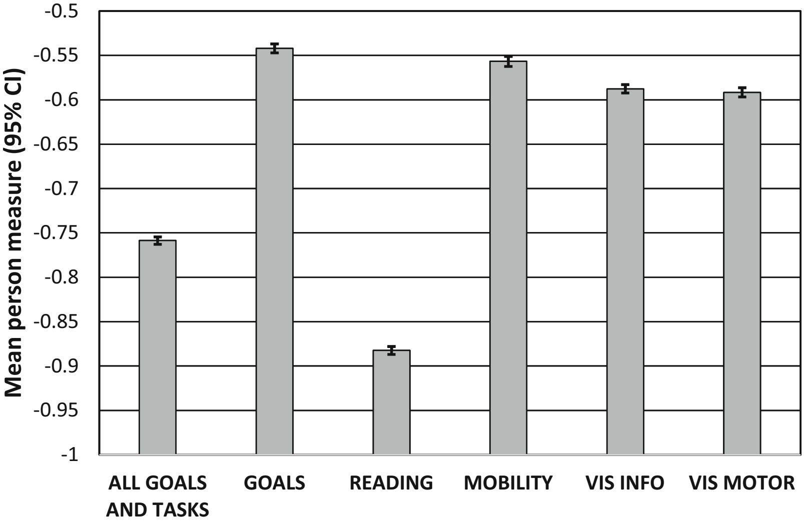
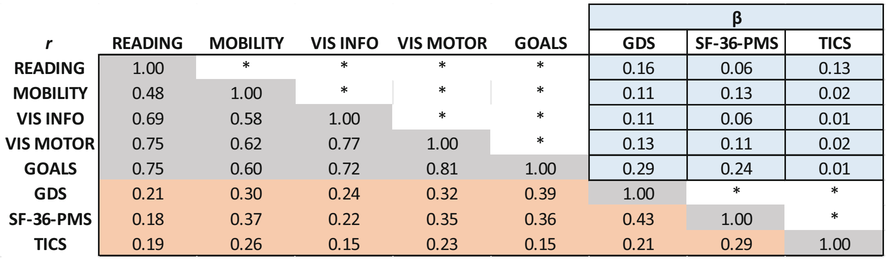


Fig. 5.5 Mean estimated person measure for each functional domain (Reading, Mobility, Visual Information Processing, and Visual Motor); mean of person measures estimated from all difficulty ratings of Goals and Tasks combined and from difficulty ratings of only Goals. Error bars represent 95% confidence intervals. All differences are statistically significant except the difference between Visual Information Processing person measures and Visual Motor person measures

Table 5.1 Correlation matrix for person measures estimated from AI difficulty ratings (gray cells) and from Rasch analysis of responses to more general health questionnaires (orange cells) – the GDS (depression), SF-36-PFS (physical ability), and TICS (cognitive disorders). The blue cells in the upper right are normalized linear regression model weights (β) on person measures estimated from GDS, SF-36-PFS, and TICS ratings to predict person measures estimated from AI difficulty ratings for each of the functional domains



						β		
r	READING	MOBILITY	VIS INFO	VIS MOTOR	GOALS	GDS	SF-36-PMS	TICS
READING	1.00	*	*	*	*	0.16	0.06	0.13
MOBILITY	0.48	1.00	*	*	*	0.11	0.13	0.02
VIS INFO	0.69	0.58	1.00	*	*	0.11	0.06	0.01
VIS MOTOR	0.75	0.62	0.77	1.00	*	0.13	0.11	0.02
GOALS	0.75	0.60	0.72	0.81	1.00	0.29	0.24	0.01
GDS	0.21	0.30	0.24	0.32	0.39	1.00	*	*
SF-36-PMS	0.18	0.37	0.22	0.35	0.36	0.43	1.00	*
TICS	0.19	0.26	0.15	0.23	0.15	0.21	0.29	1.00

variance in each measure explained by each factor), on the five sets of person measures estimated from difficulty ratings of Goals and of the four subsets of Tasks representing the functional domains.

We learned that two factors explain the correlation matrix (gray cells in Table 5.1) and account for 70% of the variance in the five sets of person measures [22, 23, 27]. The left panel of Fig. 5.6 illustrates the two orthogonal factors and the vectors for the five sets of person measures plus the vector for the principal axis (black). Reading (green) loads most heavily on factor 1 and Mobility (red) loads most heavily on factor 2. Visual information processing (yellow) and visual motor (violet) vectors, and the vector for Goals (blue) are close to the principal axis (black vector, which corresponds to all Goals and Tasks combined).

We speculate that factor 1 represents central vision (*e.g.*, altered by visual acuity and contrast sensitivity losses) and factor 2 represents peripheral vision (*e.g.*, altered by visual field loss and blind areas in vision called scotomas). Not only can these two types of visual impairment occur independently, they also have different effects on visual perception. There is neuroanatomical and neurophysiological evidence of two visual pathways in the brain following visual processing in the primary visual cortex (V1), one in the parietal cortex and the other in the temporal cortex. The parietal pathway, sometimes called the "where" system, receives most of its input from the peripheral retina and appears to be responsible for visual perceptual processing related to spatial awareness and visual control of actions, whereas the temporal pathway, sometimes called the "what" system, receives most of its visual input from the central retina and appears to responsible for object identification and interpretation of patterns [28]. Changes in visual acuity results in changes in the reading threshold (minimum size of print that can be read at all), whereas central scotomas reduce the maximum (asymptotic) reading rate that can be achieved with enlarged print [29], which suggests how the two factors can contribute independently to reduced reading function in low vision.

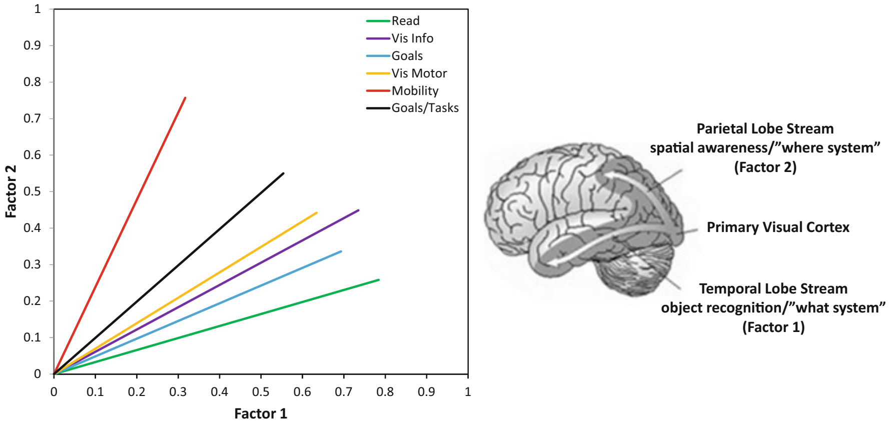


Fig. 5.6 ***Left panel***: Results of exploratory factor analysis with varimax rotation on the correlation matrix of 6 sets of person measures estimated from Task difficulty ratings in each of the 4 functional domains (green, red, purple and yellow vectors), Goal difficulty ratings (blue vector), and difficulty ratings of all Goals and Tasks combined (black vector). Vector magnitude corresponds to the square root of the variance in the person measures that is explained by the two factors and the cosine of the angle between any two vectors is equal to the correlation between the respective person measures. ***Right panel***: The Mobility vector loads most heavily on factor 2, the Reading vector loads most heavily on factor 1, and the Goals/Tasks vector loads equally on the two factors. We hypothesize that these factors are independent components of visual ability that reflect the two visual pathways observed in the parietal and temporal lobes of the cortex. To the extent that mobility relies on spatial awareness and control of actions, which are attributed to the parietal lobe that receives most of its input from peripheral retina, and reading relies on object recognition, which is attributed to the temporal lobe that receives most of its input from the central retina, we expect visual field loss and scotomas to have their largest effect on functions that depend on factor 2 and visual acuity loss to have its largest effect on functions that depend on factor 1

These two independent vision-related factors have been observed repeatedly over the past several years contributing to person measures estimated for different samples of visually impaired patients from their responses to different visual function questionaires [30, 31]. There also have been reports of significant contributions to person measure estimates from physical functioning (as measured by the SF-36 physical functioning scale [SF-36-PFS] [32]), cognitive functioning (as measured by the Telephone Interview for Cognitive Status [TICS] [33]), and psychological state (as measured by the Geriatric Depression Scale [GDS] [34], Center for Epidemiologic Studies – Depression Scale [34], or Patient Health Questionnaire-9 [35]).

However, as itemized in the last three rows of cells below the diagonal in Table 5.1 (pink highlight), the correlations (*r*) between these health state measures and AI measures for Goals and each of the four functional domains are weak [36]. These low correlations suggest that additional independent factors are required to explain the covariances that are added to visual ability estimates by co-morbidities [27, 37]. Even though individual correlations are weak, the physical, psychological, and cognitive health states are statistically significant predictors of the Goal and functional domain measures in a multivariate linear model. Table 5.1 lists above the diagonal (blue highlight) the normalized weights in the linear model (β) [36]. Although small in all cases, the GDS weight is significantly different from zero for all five functional domains; the SF-36-PFS weight is significantly different from zero for all domains except reading; and the TICS weight is significantly different from zero for all domains except visual information processing.

5.3.1 Latent Variable Model for Sources of Variance in AI Visual Ability Measures

To better understand how co-morbidities contribute to estimates of visual ability measures from AI Goal and Task difficulty ratings, we constructed the conceptual path diagram schematized in Fig. 5.7. Estimated latent intervening variables are symbolized with yellow ellipses; latent factors are symbolized with salmon and gray ellipses; and observed manifest variables are symbolized with blue rectangles. The arrows identify paths by which the inferred latent factors give rise to the observed indicators (manifest variables, which are Goal and Task difficulty ratings), both directly and by way of intervening latent variables (estimated person measures). Vision Factor 1 corresponds to Factor 1 (“what” visual processing) in Fig. 5.7 and Vision Factor 2 corresponds to Factor 2 (“where” visual processing) in Fig. 5.7.

The black arrows from the vision factors (salmon ellipses) to each intervening latent variable has a fixed weight that is estimated from the factor analysis summarized in Fig. 5.6. Each weight corresponds to the projection of the respective Goals or functional domain vector onto that factor. Each red arrow from the independent latent systemic health state factors (gray ellipses) to each intervening latent variable has a weight estimated from structural equation modeling. From the perspective of visual ability, the latent systemic health state factors are acting in the role of effect modifiers on the intervening latent variables (*i.e.*, person measures estimated from Rasch analysis). The latent visual factors also are predictors of visual impairment measures (visual acuity and visual fields by way of regression models), continuous latent variables (incorporated in Psych, Cognitive, and Physical latent intervening variables that are not shown in the paths to the manifest variables on the right in Fig. 5.7) estimated from Rasch analysis of rating scale responses to depression (GDS), cognitive functioning (TICS), and physical functioning (SF-36 PFS) questionnaires and categorical responses (scored as a dichotomous grouping variable, *i.e.*, 0,1) to a detailed health/functional/psychosocial intake history questionnaire [36].

Each observation also has sources of random variance (not shown in Fig. 5.7) and covariances (also not shown). The weights on the unfixed paths were estimated by structural equation modeling.

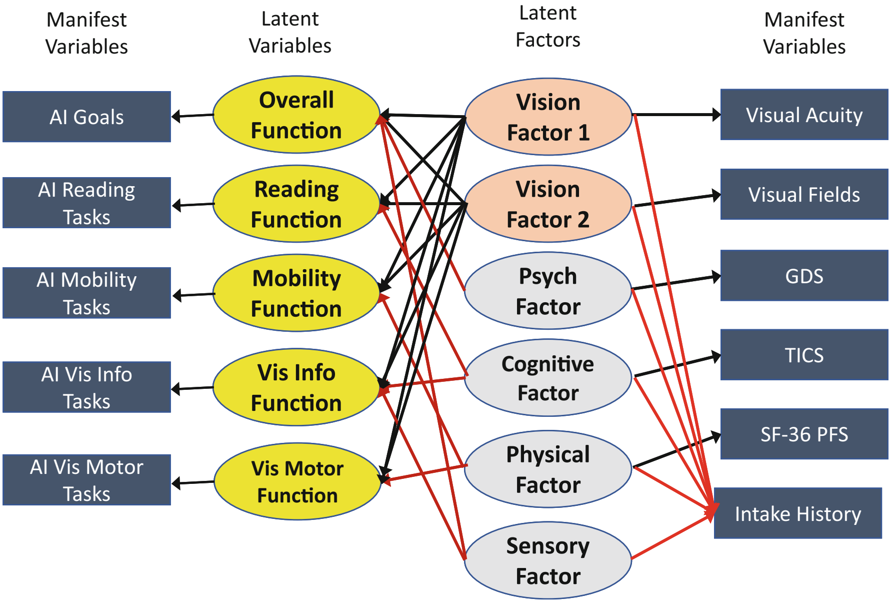


Fig. 5.7 Conceptual path model of how latent vision and non-visual factors contribute to explaining observed difficulty ratings of items in the AI (blue rectangles on the left); to observed psychophysical measures of visual impairments (blue rectangles for visual acuity and visual fields on the right); to observed psychometric measures of depression severity (GDS), cognitive impairment (TICS), and physical limitations (SF-36-PFS); and to self-reported co-morbidities (intake history). The black arrows identify paths for which weights are estimated from Rasch analysis and regression models and the red arrows identify paths for which weights are estimated from structural equation modeling

The structural equation model constructed from the conceptual design in Fig. 5.7 anchored the intervening latent visual function domain variables to person measures estimated from Rasch analysis of AI Goal and Task difficulty ratings; the latent vision factor value for each person to values estimated from principal axis factoring of the visual function domains plus visual acuity and contrast sensitivity covariance matrix; and the systemic health state factor values for each person from Rasch analysis of item responses to the GDS, TICS, and SF-36 PFS and from regression models of intake self- reported indicators in the Intake History [27]. The first row of Fig. 5.8 illustrates predicted reading function from the two vision factors vs measured reading function (left panel), predicted reading function from the 4 health state factors vs measured reading function (center panel), and predicted reading function from all 6 latent factors combined vs measured reading function (right panel). The second row makes the same comparisons for mobility function, the third for visual information function, and the fourth for visual motor function. The predicted person measures from the vision factors for each functional domain (first panel in each row) are linearly related to the measured values, but they are not accurate (i.e., they do not fall on the identity line). If health state factors make a consistent contribution to

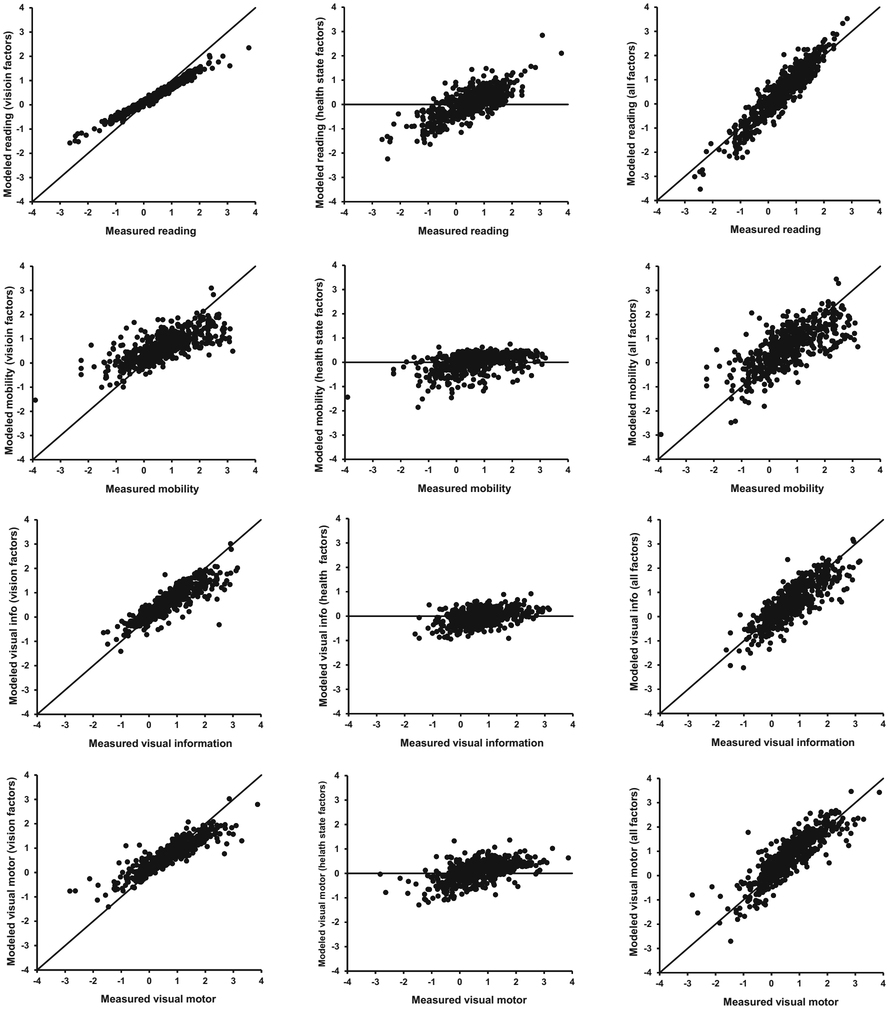


Fig. 5.8 ***Left column***: Scatter plots of person measures modeled from a general linear model (GLM) using only loadings from exploratory factor analysis (vision factor 1 and vision factor 2 in Fig. 5.7) versus the person measure estimated from Rasch analysis of difficulty ratings of AI items for each functional domain: Reading (row 1), Mobility (row 2), Visual Information Processing (row 3), and Visual Motor (row 4). The identity line is the expected relationship if only vision factors contributed to the estimated measures. ***Middle column***: Scatter plots of person measures modeled from a GLM using only loadings from the systemic factors (Psych, Cognitive, Physical, and Sensory factors in Fig. 5.7) versus person measures estimated from Rasch analysis of difficulty ratings of AI items in each functional domain. The horizontal line is the expected relationship if non-vision factors did not contribute to the measures estimated from Rasch analysis. ***Right column***: Scatter plots of person measures modeled from a GLM using loadings from all 6 factors in Fig. 5.7 versus person measures estimated from Rasch analysis of difficulty ratings of AI items in each functional domain. The identity line is the expected relationship if the 6 independent factors accounted for all contributions to the observed person measures

visual function measures, then the predicted values in the second panel of each row should correlate with the measured values. The Pearson correlations are 0.77 for Reading (first row), 0.44 for Mobility (second row), 0.44 for Visual Information (third row), 0.56 for Visual Motor (fourth row), and 0.61 for overall visual ability (estimated from Goal difficulty ratings and not shown). The predictions of measured values for the full model (2 vision factors and 4 systemic health state factors combined) are shown in the third column of Fig. 5.8. The addition of contributions from the systemic health state factors not only improve the accuracy of the predicted person measures, it also increases variance about the identity line which provides explanations of previously unexplained variance in the person measures.

5.4 Intervention-Specific Differential Item Functioning (DIF)

Upon completion of rehabilitation services, the AI is re-administered adaptively with a follow-up CATI using a slightly different algorithm. Patient ratings are elicited the same way they were at baseline, except that any Goals rated not important or not difficult at baseline are not re-administered. If at baseline a Goal was given a difficulty rating greater than not difficult, then the difficulties of that Goal's subsidiary Tasks are rated irrespective of the Goal difficulty rating at follow-up. Also, any Tasks rated not difficult or not applicable at baseline are not re-administered. The rationale for this "item-filtering" approach is that rehabilitation of Goals and Tasks rated not important or not applicable or not difficult at baseline is of no utility to the patient and those activities would not be included in the individualized rehabilitation plan.

5.4.1 Increasing Functional Reserve

In the case of vision rehabilitation, the aim of intervention is to increase the patient's ability to function, *i.e.*, increase the patient's functional reserve for activities targeted in the individualized rehabilitation plan. We typically think of increasing functional reserve by increasing the patient's visual ability, $\varphi_{nj} + \Delta\varphi_{nj} = (\alpha_n + \Delta\alpha_n) - \rho_j$, *e.g.*, improving the patient's vision by correcting refractive error with new glasses. However, we also can increase the patient's functional reserve by decreasing the visual ability required to perform the activity, $\varphi_{nj} + \Delta\varphi_{nj} = \alpha_n - (\rho_j + \Delta\rho_j)$, *e.g.*, by equipping the patient with a magnifier. Most generally, a change in the patient's functional reserve reflects either a change in the person measure and/or a change in the item measure, $\Delta\varphi_{nj} = \Delta\alpha_n - \Delta\rho_j$. Thus, the overall outcome of vision rehabilitation for person n is the average change in functional reserve over the J_n activities identified in that person's individualized rehabilitation plan:

$$\Delta\varphi_n = \sum_{j=1}^{J_n} \frac{\Delta\varphi_{nj}}{J_n} = \sum_{j=1}^{J_n} \frac{\Delta\alpha_n - \Delta\rho_j}{J_n} = \Delta\alpha_n - \sum_{j=1}^{J_n} \frac{\Delta\rho_j}{J_n} \tag{5.5}$$

If $\Delta\rho_j \neq 0$ in Eq. (5.5), we must conclude that the intervention resulted in intervention-specific DIF [21]. In this case, calibrating the AI item bank at baseline and anchoring the item measures, ρ_j, to their baseline values effectively defines $\Delta\rho_j = 0$ for all items. Thus, anchoring item measures to calibrated values forces the average change in functional reserve for person n, $\sum_{j=1}^{J_n} \frac{\Delta\varphi_{nj}}{J_n}$, into $\Delta\alpha_n$ when estimating outcome measures [38]. Although mathematically equivalent for a single patient (n), from the patient's perspective an average of changes in the difficulty experienced performing selected individual activities might not be equivalent to an equal size change in visual ability. Indeed, most low vision patients simply want to be able to "see better", not have to learn new behavior and function with the assistance of an array of activity-specific and often costly devices. The limitation of Eq. (5.5) is that $\Delta\rho_j$ has the same weight for every item even though the same size change in item difficulty for different items might have different utilities for a given person depending on their difficulty at pre-rehabilitation baseline.

5.4.2 Rehabilitation Demand and Item Filtering

If a person reported that none of the activities sampled by the AI were both important (or relevant) and difficult, then it is likely that person would have no need for rehabilitation. The visually impaired consumer's demand for rehabilitation is driven by that person's desire to be able to perform activities that he or she deems necessary to regain lost quality of life. If we use the criterion that an activity rated "not difficult" or "not important" (or not applicable) at baseline is not worthy of being included in an individualized rehabilitation plan, then that activity has no *rehabilitation demand* and the item should be dropped from the analysis [19, 20]. We refer to this selective removal of items from the analysis based on the patient's responses as "item filtering".

Filtering items by estimating the person measure at baseline from responses only to items rated at least "somewhat difficult" biases the person measure toward more negative values. The left panel of Fig. 5.9 illustrates this effect of item filtering on the person measure estimate in a scatterplot of $\alpha_{n(filtered)}$ versus $\alpha_{n(unfiltered)}$ (points) compared to the identity line (red line) for 3600 low vision patients. The person measures estimated from responses to filtered items (remaining items after removing those rated "not difficult") are more negative than or equal to person measures estimated from responses to unfiltered items (all rated items). The center panel of Fig. 5.9 shows that the difference between filtered and unfiltered person measures ($\alpha_{n(filtered)} - \alpha_{n(unfiltered)}$) is linear with a slope of -2 and an intercept of 0 as a

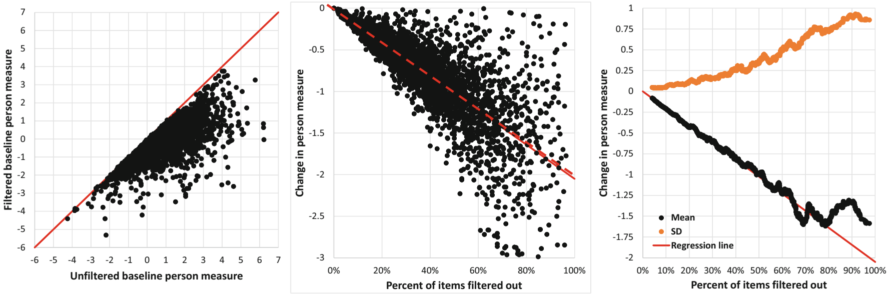


Fig. 5.9 ***Left panel***: Scatter plot of filtered person measures (*i.e.*, items rated "not difficult" excluded from the person measure estimates) versus unfiltered person measures (estimated from all item ratings). Each data point is a different person and all points are on or below the red identity line indicating that item filtering biases the person measure estimate toward lessor ability. ***Center panel***: Difference between the filtered and unfiltered person measure for each person versus the percent of items that were filtered out before the estimate was made. The red-dashed regression line fit to the data is plotted for comparison. ***Right panel***: Moving average change in person measure with item filtering (black points) is plotted along with the red regression line fit to the data in the center panel. As the average change in person measure decreases with increasing percentage of the items filtered out, the moving standard deviation of the change in person measure with item filtering increases (orange points)

function of the percent of items filtered out. Written another way, which would describe the trend in the data in the left panel of Fig. 5.9, $\alpha_{n(filtered)}$ is equal to $\alpha_{n(unfiltered)}$ plus a negative bias that is twice the percentage of items that have been filtered out for that person, $\alpha_{n(filtered)} = \alpha_{n(unfiltered)} - \frac{2\left(J_{n(unfiltered)} - J_{n(filtered)}\right)}{J_{n(unfiltered)}}$, for which $J_{n(filtered)}$ and $J_{n(unfiltered)}$ refer to the number of filtered and unfiltered items, respectively, that were rated by person n. However, note in the right panel of Fig. 5.9 that the moving standard deviation of the change in person measure increases with increasing percentage of the items filtered out (orange points) as the moving average change in person measure (black points) decreases linearly along the regression line (red line from the center panel of Fig. 5.9). The increase in variance with increases in the percentage of items filtered out is likely linked to differences between persons in the distributions of item measures among remaining items.

5.4.3 Utility to the Patient of Increasing Functional Reserve

In economics, the term *utility* refers to a quantity of how useful or desirable something is to a person. Utilities are preference values specified in relative units, often on a continuous ratio scale ranging from zero (no value) to 1.0 (maximum value of things being compared). Currently, the closest the AI comes to eliciting information from the patient about their preferences for being able to perform an activity is the elicitation of ordinal importance ratings of Goals and dichotomous responses of relevance to the individual of Tasks, both scored as 0 or 1. These binary scores are used to weight the items for the purpose of determining whether the item is filtered out or retained when estimating the person measure. Even though a polytomous scale is used to rate the importance of Goals, the importance ratings are dichotomized for the purpose of item filtering.

For an outcome measure to be truly person-centered, it must factor in the individual's preferences for specific outcomes. The term *disutility* refers to a negative utility value, *i.e.*, the utility of something a person is willing to trade to be rid of something else that is undesirable. The greater the disutility of the patient's functional state, the greater is the rehabilitation demand. In the case of vision rehabilitation, the disutility for an individual of a particular functional state should be estimated in terms of the amount of time, effort, and resources the person is willing to expend to achieve a specific less disabled state. Although we can define functional ability by how difficult it is for an individual to perform activities that are important to her or him, the utility of the specific functional ability outcome would be determined not only by the importance of the activity to the individual, but also by the level of effort required to satisfy the objective of the activity. For example, cooking daily meals may have high utility to an individual even if it is difficult to do because of its contribution to the objective of independent daily living, whereas cooking for the objective of recreation (*i.e.*, joy of cooking) may have high utility to the person only if it is easy enough to be enjoyable. Looked at another way, for the person to realize a net gain from vision rehabilitation, the utility of a functional ability outcome must be equal to or greater than the utility of the person's time, effort, and resources that must be expended to achieve that outcome.

5.4.4 Social Utility of AI Goals

The question we are raising is one of how best to measure the utility of a functional outcome. So, why not simply have the patient rate the importance of being able to perform with ease activities described by items in the AI and then use Rasch analysis of those ratings to estimate the utility of the outcome of vision rehabilitation? As part of its adaptive design, the patient already rates the importance of all the Goals in the AI. Rasch analysis assumes agreement within the population of the ordering of items, which might not be true of item importance. Applying Rasch analysis to importance ratings of AI Goals implies the existence of a consensus, which would translate to a latent variable that might best be described as the "social utility" of performing the itemized activities independently [39]. We explored this idea by using Rasch analysis to construct social utility measurement scales from importance ratings by 600 low vision patients [39].

As illustrated in Fig. 5.10, we observed that the putative social utility of performing personal hygiene activities independently (item measure = 3.39) is greater than the social utility placed on shopping independently (item measure = 1.86), but this is not necessarily true for an individual. The level of consensus

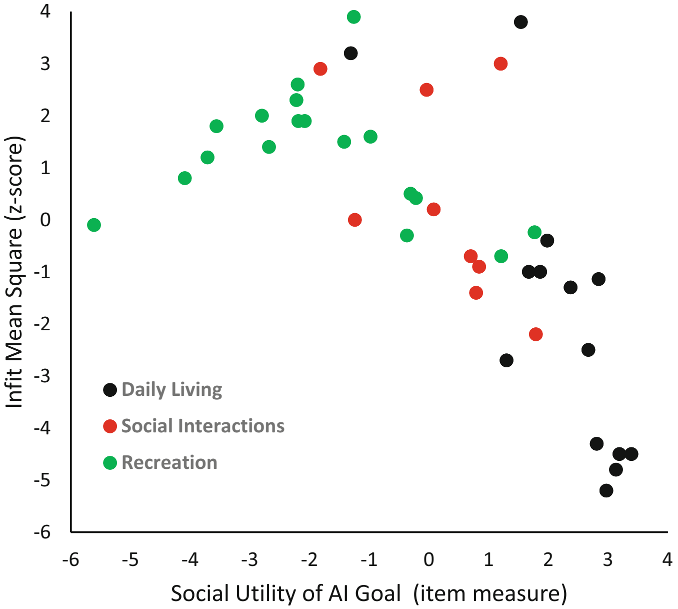


Fig. 5.10 Scatter plot of the z-score for the item infit mean square versus the item measure estimated from Rasch analysis of importance ratings of AI Goals. Each point is 1 of the 50 AI Goals color-coded by the Objective in the Activity Breakdown Structure (Fig. 5.1) to which the Goal is assigned: Daily Living (black), Social Interactions (red), or Recreation (green). The infit z-score is 0 if variance of the distribution of deviates from the item measure estimate is at the expected value, negative if the variance is less than expected (high agreement between persons), or positive if the variance is more than expected (high disagreement between persons)

(dispersion between people of importance ratings relative to the expected ratings for each Goal) is captured in the Rasch model infit statistic and, as shown in Fig. 5.10, consensus is strong (low dispersion) for AI Goals with high social utility (Daily Living – black circles) and Goals with low social utility (Recreation – red circles) and is weakest (high dispersion) for AI Goals with medium social utility (Social Interactions – green circles), giving rise to a parabolic relationship between consensus and estimated social utility (*i.e.*, people agree most on how to order the importance of Goals on high valued activities and on low valued activities). Overall, AI Goals under the Daily Living Objective have the highest social utility (mean item measure = 2.17; SD = 1.21), Goals under the Social Interaction Objective have intermediate social utility (mean = 0.26; SD = 1.16), and Goals under the Recreation Objective have low social utility (mean = −1.82; SD = 1.16).

The objective of vision rehabilitation is to make activity goals that are important to the individual and difficult or impossible to achieve easier by way of vision enhancement methods and technology, adaptations (*i.e.*, adopting new strategies and using assistive technology that obviate performance of customary vision-dependent tasks), and modified independence (limited use of human assistance to overcome intractable or high safety risk barriers that prevent attainment of a specific end goal). In terms of the theory behind the AI, the amount of difficulty a person experiences when attempting a specific activity is determined by functional reserve (φ_{nj}). The objective of vision rehabilitation is to increase functional reserve, thereby reducing the difficulty of performing the activity, by modifying the way activities that are important to the person are performed.

Although the concept of social utility might be useful for policy making, it is not a good starting point for developing a truly person-centered rehabilitation plan and a person-centered measure of visual function outcomes in terms of net gain to the individual. To achieve this aim, we must start with a model of utility measures of vision rehabilitation functional outcomes in terms of reductions in rehabilitation demand and verify the internal validity of the model and analytic methods by applying them to simulations of the approach the model implies. The change in difficulty of performing different activities that are important and difficult to the patient can be estimated by the Rasch model from the change in rehabilitation demand, which is equal to the change in functional reserve, $\Delta\varphi_n = \Delta\alpha_n$ in Eq. (5.3) when item measures are anchored to baseline values so $\Delta\rho_j = 0$ for all items and all changes are forced into the person measure. However, the utility of the functional outcome for each specific activity is likely to be idiosyncratic to an individual patient. For example, the utility of a small realized change in functional reserve for cooking might have high value when serving the Daily Living Objective (social utility of preparing daily meals is 1.30), less value when serving the Recreation Objective (social value of recreational cooking and baking is −0.98), and even less value when serving the Social Interaction Objective (social utility of cooking to entertain guess is −1,24).

5.4.5 Utility of Vision Rehabilitation Outcomes

The utility of reducing rehabilitation demand of item j for person n to zero is υ_{nj}. Since we are referring to a single item, upsilon denotes a marginal utility, which is a person-specific function of both the item's difficulty to the person (D_{nj}) and the importance to the person of being able to perform the activity without difficulty (I_{nj}). While φ_{nj}, which determines the ordinal value of D_{nj}, is a continuous latent variable, no equivalent continuous variable has been modeled for determining the ordinal importance rating, I_{nj}. Therefore, to illustrate the derivation of a utility weighting model for the AI, we will employ discrete ordinal variables assigned by person n to the difficulty rating (D_{nj}) and importance rating (I_{nj}) of AI Goal j. The marginal utility of totally successful vision rehabilitation (*i.e.*, achieve a non-disabled state) of Goal j for person n is $\upsilon_{nj} = U_n(I_{nj}, D_{nj})$.

To paraphrase Gertrude Stein, the utility of a utility is a utility. By definition, we assume that the utility of vision rehabilitation on Goal j for the average person, $U(I_j, D_j)$, is a function of the respective part worth utilities associated with the importance and difficulty ratings, conditioned on the kth Objective, $O_{k(j)}$, so $U(I_j, D_j) = f(U_i(I_j|$

$O_{k(j)}$), $U_d(D_j| O_{k(j)})$). Individuals will randomly deviate from this average relationship, so we model the mapping of I_{nj} and D_{nj} to the utility of vision rehabilitation for Goal j for person n as $U_n(I_{nj}, D_{nj}) = U(I_{nj}, D_{nj}) + \epsilon_{nj}$, where (I_{nj}, D_{nj}) are the ratings of Goal j by person n and ϵ_{nj} is the randomly distributed deviate. If either the importance or difficulty of Goal j is 0 for person n, then the utility of vision rehabilitation for Goal j will be zero, i.e., $U(I_{nj}, D_{nj}) = f(U_i(0| O_{k(j)}), U_d(D_{nj}| O_{k(j)})) = f(U_i(I_{nj}| O_{k(j)}), U_d(0| O_{k(j)})) = 0$, and by definition $\epsilon_{nj} = 0$ when $I_{nj} = 0$ or $D_{nj} = 0$.

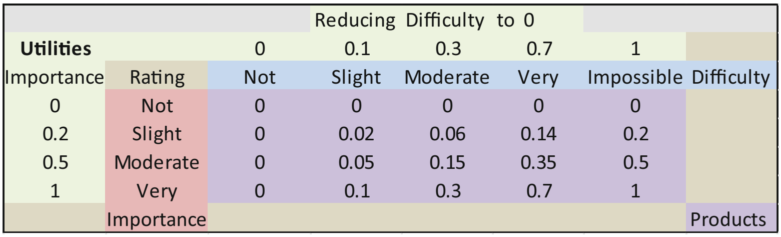
We are conditioning the utility function for each Goal on the Objective it serves because of the observed segregation by Objective of social values (estimated from Goal importance ratings) and level of consensus in the LV population, which is summarized in Fig. 5.9. Conditioning by Objective also is motivated by the possibility that the relation of Goal utility to difficulty may be determined by the reason for doing the activity (e.g., cooking to maintain independent living vs joy of cooking). If there are three different utility functions, one for each of the three Objectives, then we expect differences in utilities between Goals to be represented by the distance between them in a 3-dimensional utility space.

The marginal utility of Goal j is likely to be characterized by a nonlinear function (unique to Objective k) of the part worth utilities associated with the observed ordinal importance and difficulty ratings of the Goal. For purposes of estimation, we can approximate this nonlinear function with a Taylor series:

$$U\left(I_{nj}, D_{nj}\right) \approx b_0 + b_1 I_{nj} + b_2 D_{nj} + b_3 I_{nj}^2 + b_4 D_{nj}^2 + b_5 I_{nj} D_{nj} + \cdots$$

in which b_0 is a constant and the other coefficients correspond to factorial-weighted first, second, and higher order partial derivatives (ellipsis denotes the higher order terms in the infinite series). To illustrate how this model can be applied to estimate utilities of vision rehabilitation outcomes, we simulated the method. The simplifying assumptions made within the model to create the simulation are: (1) the utility function is the product of the part worth utilities for importance and difficulty, *i.e.*, $U(I_{nj}, D_{nj}) = U_i(I_{nj}) \times U_d(D_{nj})$, which yields an overall utility of zero if either part worth utility is zero, an overall utility of 1 if both part worth utilities are 1, and an overall utility less than or equal to the lowest part worth utility; (2) the utility function is the same for all objectives, i.e., $U_{k(j)}(I_{nj}, D_{nj}) = U(I_{nj}, D_{nj})$ for all k; and (3) the relation of utility to importance/difficulty rating combinations is fixed for the population, *i.e.*, $\epsilon_{nj} = 0$. These assumptions enable us to estimate a variable that is equivalent to utility by constructing a dissimilarity matrix comparing all possible deterministic importance/difficulty pairs (*I,D*).

In the AI, ordinal values for importance ratings range from 0 to 3 (where 0 is "not important" and 3 is "very important") and ordinal values for difficulty ratings range from 0 to 4 (where 0 is "not difficult" and 4 is "impossible"). As schematized in Fig. 5.11, we already have defined the utility of LVR for Goals with ratings (*I,0*) or (*0,D*) to be zero for all values of *I* and *D*, leaving 12 (*I,D*) combinations for which the partial utility associated with Goal importance and decrease of difficulty are greater than zero. For the simulation we simply filled in arbitrarily-chosen ascending values



			Reducing Difficulty to 0				
Utilities		0	0.1	0.3	0.7	1	
Importance	Rating	Not	Slight	Moderate	Very	Impossible	Difficulty
0	Not	0	0	0	0	0	
0.2	Slight	0	0.02	0.06	0.14	0.2	
0.5	Moderate	0	0.05	0.15	0.35	0.5	
1	Very	0	0.1	0.3	0.7	1	
	Importance						Products

Fig. 5.11 Relation of utilities to AI Goal importance and difficulty ratings as used in the simulation to test estimation of the utility to the patient of vision rehabilitation outcomes. The 4 Goal importance ratings range from "Not Important" to "Very Important" and are assigned the part worth utilities in the left most column. The 5 Goal difficulty ratings range from "Not difficult" to "Impossible" and are assigned the part worth utilities (of reducing difficulty to 0) in the top row. The marginal utility for each Goal is the product of the part-worth utilities corresponding to Importance and Difficulty ratings assigned by the patient to the Goal (entries in the purple cells)

for the marginal utilities (light green areas) and products of the marginal values for each (*I,D*) combination (purple area). Thus, a triangular matrix for the AI has 66 unique paired comparisons of different non-zero (*I,D*) combinations.

We can think of the marginal vision rehabilitation utility of Goal *x* with importance and difficulty ratings (I_x,D_x) as mapping to a point U_x on a number line that represents the utility of time, effort, and resources that would have to be expended on vision rehabilitation. Similarly the partial vision rehabilitation utilities of Goals *y* and *z* with respective importance and difficulty ratings (I_y,D_y) and (I_z,D_z), would map to different points, U_y and U_z, on the same number line for the expenditure of time, effort, and resources. Individuals then can be asked to judge the relative distances between each pair of points.

For example, the person would be asked, "In terms of allocating your time, effort and resources to rehabilitation, which of the three Goals would you give the highest priority?" That question would be followed by, "Which of the remaining two Goals would you give the lowest priority?" The final question would be, "Is the priority you give to the left-over Goal closer to the highest priority Goal or the lowest priority Goal?" Let's imagine the person gave the highest priority to Goal *x*, the lowest priority to Goal *y*, and said that the priority of Goal *z* is closer to the priority of Goal *y*. Thus, on the utility number line, the distance between U_x and U_y is the largest of the three comparisons and the pairing of (I_x,D_x) and (I_y,D_y) in the matrix would be assigned a dissimilarity rank of 3; the distance between U_y and U_z is the smallest of the three comparisons and the pairing of (I_y,D_y) and (I_z,D_z) in the matrix would be assigned a dissimilarity rank of 1; and the remaining pairing of (I_x,D_x) and (I_z,D_z) corresponds to an intermediate distance and would be assigned a dissimilarity rank of 2 in the matrix.

Repeating these judgments and assignment of dissimilarity rank scores for all feasible triadic comparisons across all persons in the sample, and averaging relative distance rank scores in each cell of the triangular dissimilarity matrix, we can employ non-metric unidimensional scaling [40] (UDS) to map each (*I,D*) rating pair to a variable that is monotonic with vision rehabilitation utility.

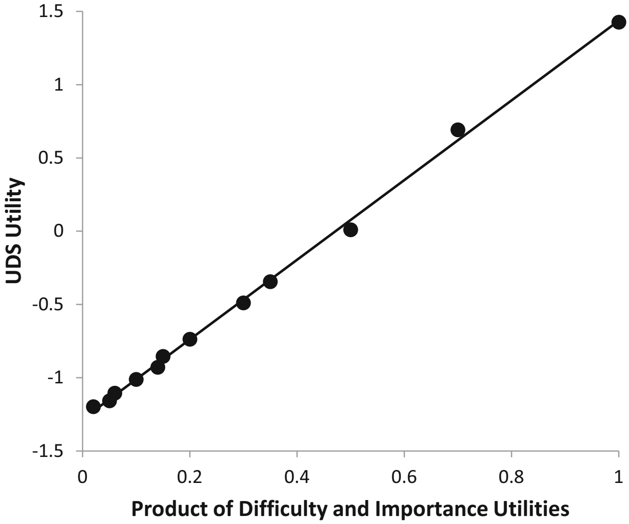


Fig. 5.12 Utilities estimated from unidimensional scaling (UDS) of sums of ordinally-ranked differences in utility of 3 unique combinations of Goal difficulty and importance ratings (*i.e.*, triadic comparisons) as a function of the product of corresponding part worth utilities for the ratings. In this case, because the monotonic function was exponential, the log of the estimated distance metric is plotted as a function of the 12 marginal utilities defined in Fig. 5.11 and the linear relationship defines log distance as the UDS-estimated utility

Using the table in Fig. 5.11, we simulated the triadic comparison judgments by assigning rank scores to (*I,D*) pairings in all 1320 possible triads based on the products of the marginal utilities assigned to importance (*I*) and difficulty (*D*) resulting in 20 dissimilarity ranks contributing to the average in each of the 66 cells of the triangular matrix. [*N.B.* This scaling method effectively made $\epsilon_{nj} = 0$ for all simulated judgments]. Figure 5.12 illustrates a scatter plot comparing the estimated utility (in arbitrary units) for each of the 12 (*I,D*) pairs from unidimensional scaling (UDS) versus the products of the partial utilities assigned to importance and difficulty ratings in each pair (values in purple area of Fig. 5.11), which were used to make the ranked distance judgments in each triadic comparison.

When applied to the real world, we will not know the true utilities associated with the importance and difficulty rating categories as we do for this simulation. The UDS (and more generally multidimensional scaling [MDS], which would be used if the Objectives represent different utility dimensions) estimates distances using the ordinal data in the dissimilarity matrix (non-metric scaling). The UDS (and MDS) approach enables us to perform statistical tests of the goodness of fit of the estimated distances in the dissimilarity matrix to the average ranks of dissimilarity scores from the triadic comparisons (*e.g.*, Shepard plots and estimates of "stress").

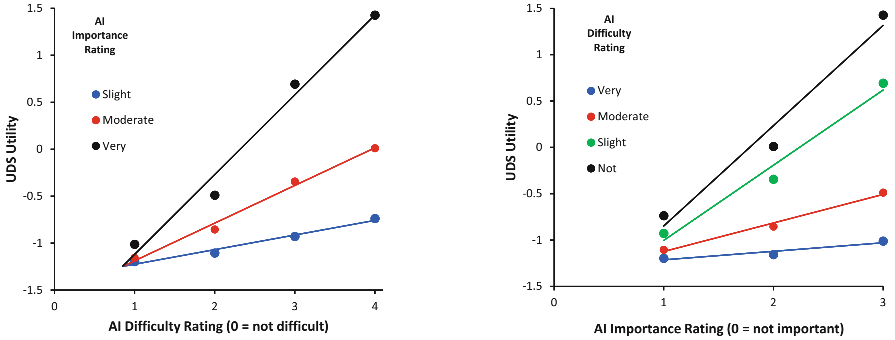


Fig. 5.13 ***Left panel***: Estimated utilities as a function of the difficulty ratings of AI Goals for each rating of importance. The slope of the line fit to the results is a function of the importance ratings. ***Right panel***: Estimated utilities as a function of the importance ratings of AI Goals for each rating of difficulty. The slope of the line fit to the results is a function of the difficulty ratings

The left panel of Fig. 5.13 illustrates a scatterplot of marginal utility estimates of $U(I,D)$ from UDS coordinates vs ordinal ranks of different non-zero difficulty rating categories for the 3 different levels of non-zero importance ratings (color-coded). These linear relationships in the left panel of Fig. 5.13 imply $U(I,D) = m_I D$, for which the slope m_I is dependent on I. The right panel of Fig. 5.13 similarly illustrates a scatterplot of UDS estimates of $U(I,D)$ vs ordinal ranks for different non-zero importance ratings for the 4 different levels of non-zero difficulty ratings. The linear relationships in the right panel of Fig. 5.13 imply $U(I,D) = m_D I$, for which the slope m_D is dependent on D. The abscissa value where the lines converge corresponds to the location of zero difficulty or zero importance, respectively (ordinal rank scores arbitrarily are equally spaced on the abscissa). The horizontal deviations of the points from the line must be construed as errors resulting from the assumption that the ordinal ratings represent equal intervals. From these linear relationships, we conclude that $\frac{\partial U(I,D)}{\partial I} = m_D$ and $\frac{\partial U(I,D)}{\partial D} = m_I$. In Fig. 5.11 we defined the overall utilities estimated from UDS to be the product of the respective importance and difficulty partial utilities, $U(I,D) = U_i(I) \times U_d(D)$. This definition means that the Taylor series approximation is first-order, $\frac{\partial U(I,D)}{\partial I} = \frac{\partial U_i(I)}{\partial I} U_d(D)$ and $\frac{\partial U(I,D)}{\partial D} = \frac{\partial U_d(D)}{\partial D} U_i(I)$. Combined with the conclusions we drew above from the linear relationships in Fig. 5.13, we can see that the estimated slopes must be linear with the respective partial utilities, $m_D = \frac{\partial U_i(I)}{\partial I} U_d(D)$ (blue points in Fig. 5.14) and $m_I = \frac{\partial U_d(D)}{\partial D} U_i(I)$ (red points in Fig. 5.14). Figure 5.14 confirms the tautology, which validates the analysis – we can estimate overall utilities from UDS (or more generally, MDS) on a matrix of average dissimilarity rank scores obtained from triadic comparisons using the pre-assigned overall utilities to rank distances.

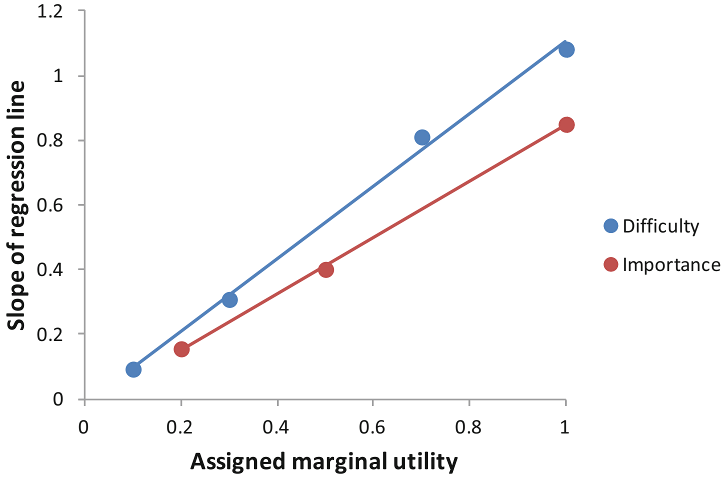


Fig. 5.14 Slopes of the regression lines in Fig. 5.13 as a function of the marginal utility corresponding to the assigned part-worth utilities for each combination of non-zero ordinal importance (red) and difficulty (blue) ratings of AI Goals

5.4.6 *Extension of the Utility Model to Estimation of Net Gain from Vision Rehabilitation*

Consider the functional outcome for person *A* who has 3 Goals in her individualized rehabilitation plan and the functional outcome for person *B* who has 10 Goals in his individualized rehabilitation plan, for which 3 of them are the same as the Goals in person *A*'s plan with the same combinations of importance and difficulty ratings. If vision rehabilitation results in the difficulty of all 3 of *A*'s Goals being reduced to zero and the same 3 of *B*'s Goals also reduced to zero, but less than full reduction in difficulty for some of *B*'s other 7 Goals, how would the net gain from vision rehabilitation for patient *A* compare to the net gain for patient *B*? Using the current method of measuring functional outcomes from vision rehabilitation, which is equivalent to the average change in functional reserve for Goals in the individualized rehabilitation plan, the effect size for *B* could be larger than the effect size for *A*, or vice versa, depending on the magnitudes of changes in functional reserve for *B*'s other 7 Goals relative to the change in the 3 that are the same as *A*'s.

However, from another perspective, *A* would have no need for additional rehabilitation after completion of vision rehabilitation, whereas *B* would. So, for every scenario short of all 10 of *B*'s Goals being reduced to zero difficulty, in terms of remaining *rehabilitation demand* [19], *A*'s functional outcomes would have greater utility than *B*'s. If the difficulty of all 10 of *B*'s Goals were reduced to zero, then neither *A* nor *B* would have any remaining rehabilitation demand and the utility of additional vision rehabilitation would be zero for both. The conundrums raised by this example suggest that to be truly person-centered, we should define the utility of the vision rehabilitation function outcome as the difference between the disutilities (rehabilitation demands) of post-rehabilitation and pre-rehabilitation. The question we must answer is, how do we combine the marginal utilities of reducing rehabilitation demand for different Goals to estimate the multi-attribute (all Goals combined) utility of reducing an individual's overall disability?

In a manifestation of dynamics related to the law of diminishing marginal utility, *B* might temporarily be euphoric with the reduction of difficulty of the 3 Goals having high rehabilitation demand, only to have the emotional high of successful vision rehabilitation dissipate and the disutility of the remaining rehabilitation demand emerge. The multi-attribute disutility function we seek, which combines disutilities of rehabilitation demand for all Goals in the patient's individualized rehabilitation plan, could be the linear sum of disutilities for individual Goals, or a nonlinear combination that ranges from diminishing rate of change in marginal utility (attenuation effect) to increasing rate of change in marginal utility (amplification effect) with increasing numbers of Goals having non-zero rehabilitation demand. This range of options can be expressed with a Minkowski distance, $\upsilon_{nJ_n} = \left(\sum_{j}^{J_n} u_{nj}^{b} \right)^{1/b}$ where υ_{nJ_n} is the multi-attribute utility of vision rehabilitation (or multi-attribute disutility of rehabilitation demand) for an individualized rehabilitation plan for patient n with J_n Goals after filtering.

Multi-attribute utility is a linear sum for $b = 1$; attenuation corresponds to $b > 1$; and amplification corresponds to $b < 1$. The Minkowski distance (ordinate) across Goals with the same marginal utility ranging from 0.05 to 1.0 (colored functions) are shown in Fig. 5.15 as a function of the number of Goals ranging from 1 to 10 (abscissa) in the patient's plan. The left panel of Fig. 5.15 depicts attenuation in the growth of multi-attribute utility with $b = 2$; the middle panel of Fig. 5.15 shows the linear sum in the growth with $b = 1$; and the right panel of Fig. 5.15 shows amplification in the growth with $b = 0.6$.

5.5 Visual Ability Outcomes of Vision Rehabilitation

In the preceding section we reviewed a strategy for measuring visual ability outcomes in terms of net gain to the patient by way of a multi-attribute utility model. This model entails summed changes in the utilities of reducing the difficulty (and/or reducing the importance) of attaining individual AI Goals. However, at the current stage of development, operations in the multi-attribute utility model employ ordinal rank scores assigned to the person's importance and difficulty ratings of being able to perform activities. Ideally, we would define utility to be a function of continuous latent variables for importance and difficulty, $\upsilon_{nj} = u_n(\iota_{nj}, \delta_{nj})$. Difficulty is the inverse of functional reserve, $\delta_{nj} = -\varphi_{nj} = \rho_j - \alpha_n$, a continuous latent variable

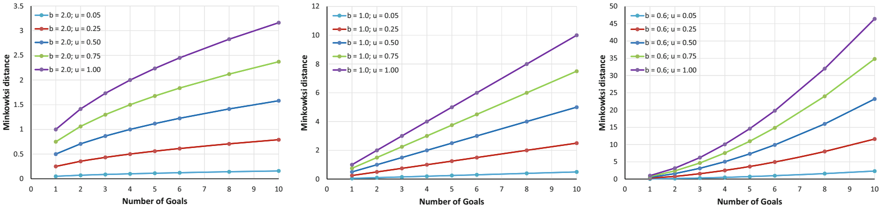


Fig. 5.15 ***Left panel***: Multi-attribute utility as a function of the number of Goals, estimated from marginal utilities using a Minkowski distance formula with the exponent variable $b = 2$, which results in an attenuation in the rate of growth with an increasing number of Goals. The marginal utility is the same for every Goal with a different value for each curve ranging from $u = 0.05$ *to* 1.0 (see legend). ***Center panel***: Same as the left panel but for a Minkowsi distance formula in which the exponent variable $b = 1$, which results in a constant rate of growth with an increasing number of Goals. ***Right panel***: Same as the left and center panels but for a Minkowsi distance formula in which the exponent variable $b = 0.6$, which results in an acceleration of the rate of growth with an increasing number of Goals

we already know how to estimate. However, estimating ι_{nj}, which is the strength of the personal preference assigned to item j by person n, is a thornier problem because of the lack of consensus between people in the ordering of items by personal preferences (*cf.*, "social utility").

We will return to the issue of how the multi-attribute utility model can be used to estimate the net gain from vision rehabilitation. But first we must review two methods of measuring patient-centered visual ability outcomes of vision rehabilitation with the AI in terms of (1) the distribution of changes in a continuous visual ability outcome variable and (2) the likelihood of attaining a change in a visual ability clinical endpoint. In both cases we employ the average change in functional reserve as the measure of the patient's self-reported visual ability before and after rehabilitation.

5.5.1 Continuous Visual Ability Outcome Measure: Average Change in Functional Reserve

The Low Vision Depression Prevention Trial in Age-Related Macular Degeneration (VITAL) was a randomized, attention-controlled, clinical trial to determine the effectiveness of behavior activation therapy as a supplement to in-home vision rehabilitation with an occupational therapist in preventing the development of major or minor depression in low vision patients with subsyndromal depressive symptoms [41]. Low vision patients were randomized to six weekly sessions of vision rehabilitation provided in the home by an occupational therapist who also provided behavior activation therapy (BA – the treatment being tested) or to six weekly sessions of supportive therapy (ST – a placebo attention control) provided in the home by a clinical social worker. The ST control group received no additional low vision services. The primary outcome measure, which was administered prior to any low vision services (PRE) and again at 2 months after the completion of services and assigned therapy (POST), was the PHQ-9, which was used to determine if the patient exhibited depressive symptoms consistent with DSM-IV criteria for major or minor depression. The AI also was administered PRE and POST low vision services and psychotherapy. Prior to randomization, all study participants received standard optometric low vision consultations; required vision assistive equipment was dispensed to all participants at study expense; and all participants were trained at the low vision clinic in how to use the equipment.

Rasch analysis of the PHQ-9 responses was used to estimate person measures of depression severity PRE and POST low vision services [35]. Rasch analysis also was used to estimate overall visual ability from AI Goal difficulty ratings with anchored item measures and thresholds for both the BA treatment group and ST control group (with item filtering) prior to receiving low vision services and again 2 months after completion of low vision services. The BA treatment group exhibited a statistically significant improvement in visual ability (Cohen's d $= 0.71$; p < 0.001). The ST control group also exhibited a statistically significant improvement in visual ability (Cohen's d $= 0.55$; p $= 0.003$). However, the distributions of change in visual ability for the BA treatment group was not significantly different from the change for the ST control group (Cohen's d $= 0.10$; p $= 0.39$) [35]. The significant medium size effect (POST-PRE) seen for both groups most likely can be attributed to the low vision devices and services that were provided in the clinic after the PRE measures of visual ability, but before the in-home vision rehabilitation supplemented by BA psychotherapy for the treatment group or the sham psychotherapy with no additional vision rehabilitation for the control group.

Although there was no difference between groups for the effects of study intervention on visual ability, the primary outcome for the VITAL study was a psychiatric clinical diagnosis in low vision patients of major or minor depression as defined by a criterion PHQ-9 score. As a study eligibility criterion, none of the participants in the study had a PHQ-9 score at PRE that exceeded the threshold for a clinical depression diagnosis. Thus, the PHQ-9 threshold for depression defined a clinical endpoint, which was exceeded at POST by significantly more patients in the ST control group than in the BA treatment group. The take-home conclusion of the VITAL study was that in-home vision rehabilitation supplemented with BA psychotherapy prevented the development of clinical depression in at-risk low vision patients.

Applying Rasch analysis to PHQ-9 item responses results in valid estimates of interval-scaled continuous person measures that can be interpreted as *depression severity* [35]. The left panel of Fig. 5.16 shows PRE (blue curve) and POST (red curve) cumulative frequency functions of PHQ-9 person measures for patients assigned to the ST control group. Negative person measures correspond to low depression severity and positive person measures correspond to high depression severity.

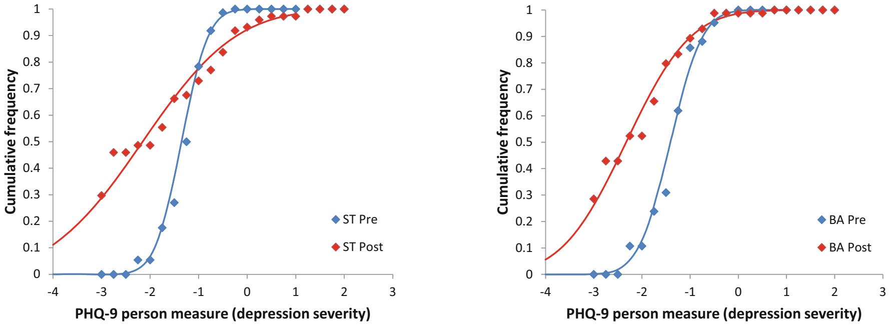


Fig. 5.16 ***Left panel***: Relative cumulative frequency of PHQ-9 person measure values of depression severity at baseline (blue curve) and at post-intervention follow-up (red curve) for the supportive therapy control group in the VITAL study. The shift of the median value to the left on the person measure axis at follow-up indicates a decrease in depression severity post-intervention. However, the curves cross, which means that about 80% of patients in the ST group had a decrease in depression severity after intervention, whereas about 20% of patients had an increase – the main effect reported for the study. ***Right panel***: Same results for the BA group as shown for the ST group at baseline, but at follow-up the cumulative functions do not cross, which indicates that nearly all patients in the BA group had a decrease or no change in depression severity at follow-up

Similarly, the right panel of Fig. 5.16 shows Pre and Post cumulative frequency functions of PHQ-9 person measures for patients assigned to the BA treatment group. The decrease in depression severity from PRE to POST (leftward shift of the red curve relative to the blue curve) demonstrates a large significant effect of intervention for both the BA treatment group (Cohen's d = 2.12; $p < 0.001$) and for the ST control group (Cohen's d = 2.02; $p < 0.001$). There is no difference between depression severity distributions for the two groups at baseline (t-test; $p = 0.28$ for PRE).

There also is no difference between the means of the depression severity distributions for the two groups at follow-up (t-test; $p = 0.25$ for POST), but the slope of the depression severity cumulative distribution is shallower for the ST group than for the BA group. This change of slope that causes the ST PRE and POST curves to cross in the right panel of Fig. 5.16 underlies the main effect of BA psychotherapy supplementing in-home vision rehabilitation preventing the development of major or minor depression, which was reported as the VITAL study primary outcome. But this study also shows that the low vision devices and services provided in the clinic, weekly in-home sessions with a professional therapist or counselor, and whatever else the two groups have in common result in a large reduction of severity in depression symptoms. Since the two groups in the study had equivalent improvements in visual ability, we explored changes in that variable as a possible explanation.

5.5.2 *Minimum Clinically Important Difference in Visual Ability as a Clinical Endpoint*

With reference to Eq. (5.5), anchoring AI item measures and thresholds to baseline values forces all intervention-specific DIF into changes in the person measure, $\Delta\alpha_n$, with all randomly distributed measurement error incorporated in the person measure estimates. The lower bound on the standard error of the person measure estimate for person n (SE_n) is proportional to the standard deviation on ζ in expression (5.1) and inversely proportional to the square root of the number of items rated by person n. The left panel of Fig. 5.17 displays the distribution of standard errors of the person measure estimate at baseline versus the person measure for VITAL study participants. To estimate the standard deviation of the person measure error distribution, we multiplied each standard error of the person measure estimate by the square root of the number of items rated by the person. The standard deviations of the person measure error distributions, so estimated for each patient, are plotted as a function of the patient's estimated visual ability (person measure) in the middle panel of Fig. 5.17.

However, the conventional logistic Rasch model normalizes the estimated measures to the standard deviation of ζ, so the expected value of the standard deviation of the error distribution should be 1 for each person. The U-shaped functional relationship between the estimate of SD_n and α_n in the middle panel of Fig. 5.17 (which shows that all SD_n values are greater than 1) can be attributed to the increased uncertainty at the extremes of the person measure distribution due to the progressive change in frequency of responding with the half-open categories ("not difficult" or "impossible" that extend from τ_4 to ∞ and from τ_1 to $-\infty$, respectively). The right

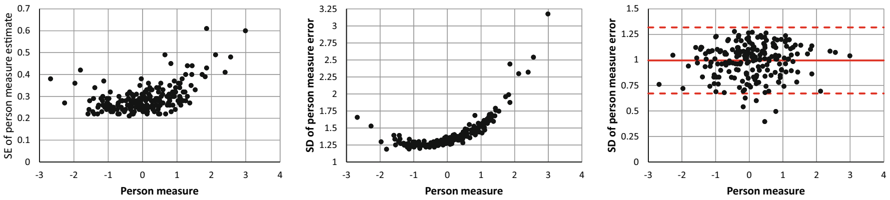


Fig. 5.17 ***Left panel***: Scatter plot of the standard error (SE) of the person measure estimates from difficulty ratings of AI Goals versus the estimated person measure at baseline in the VITAL study. ***Center panel***: The same results shown in the left panel but with each SE value multiplied by the square root of the number of items rated by that person resulting in an estimate of the standard deviation of the distribution of deviates (ζ). Notice the U-shape in the plot of the data which can be attributed to greater numbers of difficulty ratings corresponding to half-open intervals contributing to the estimate as person measures become more extreme. Also note that all values are greater than the expected value of 1. ***Right panel***: Person measure standard errors were re-estimated after omitting items for which the response represented a half-open interval. The revised standard error estimates were multiplied by the square root of the number of items retained in the estimate (points). These revised estimates of the standard deviation of the deviates, ζ, have an average value of 1 (red line) – the expected value. The red dashed lines define the 95% confidence interval

panel of Fig. 5.17 displays the distribution of SD_n when the value is estimated by multiplying SE_n by the square root of the number of items the person rated with non-extreme difficulty categories (i.e., "somewhat difficult", "moderately difficult", and/or "very difficult"). The average estimated standard deviation of ζ is 1 (solid red line), which agrees with the measurement scale normalization built into the model. The dashed lines in the right panel of Fig. 5.17 bound plus and minus two standard deviations of the between person distribution of SD_n estimates.

The variance of each person's error distribution, $\sigma^2_{\zeta_{n,j,x}}$, is the sum of within and between person squared deviations from the expected value of zero, $\mathbb{E}\left\{\zeta^2_{n,j,x}\right\}$, whereas the variance of the between person error distribution is $\sigma^2_{j,x} = \sigma^2_j + \sigma^2_x + 2r_{j,x}\sigma_j\sigma_x$, as introduced earlier in this chapter, in which σ^2_jrefers to between person variance in the item measure for the jth item and σ^2_x refers to between person variance in the threshold for the xth response category. All within person variance can be assigned to visual ability, $\sigma^2_{\alpha_n}$, so the total variance of each person's measurement error distribution is $\sigma^2_{\zeta_{n,j,x}} = \sigma^2_{\alpha_n} + \sum_{j=1}^{J_n} \sigma^2_{j,x}$. But, in the case of comparing PRE to POST intervention measures for each person, the item measures, ρ_j, and response category thresholds, τ_x, are fixed to calibrated values that are the same for both measures (the deviates due to between person differences are fixed and manifest as the same person-dependent bias for PRE and POST measures, so $\sum_{j=1}^{J_n} \sigma^2_{j,x} = 0$) and the error variance on the estimate of $\Delta\alpha_n$ is determined entirely by within person variance, $2 \times \sigma^2_{\alpha_n}$. In most cases, the number of items rated (J_n) is the same at PRE and POST, however, that is not a requirement. Also the proportion of items rated with extreme response categories is likely to be different between PRE and POST, which will differentially affect the standard error of the estimate, even when there is no change between PRE and POST in within person variance. Thus, the standard error of the estimate of $\Delta\alpha_n$ is $SE_{\Delta\alpha_n} = \sqrt{SE^2_n(Pre) + SE^2_n(Post)}$.

The smallest change in the person measure of an individual that we can say with confidence represents a real change in response to an intervention is called the *minimum clinically important difference* (MCID). The MCID is a clinical endpoint. We transform the clinical outcome for person n to a t-statistic, $t(\Delta\alpha_n, df_n) = \frac{\Delta\alpha_n}{SE_{\Delta\alpha_n}}$ with $df_n = J_n(Pre) + J_n(Post) - 2$, and the MCID for person n as the t value that corresponds to a criterion probability of making a type I error (*e.g.*, $p = 0.05$). If $t(\Delta\alpha_n, df_n)$ exceeds the criterion corresponding to the chosen p value, then MCID = 1 for person n, otherwise MCID = 0.

The odds of MCID = 1 is 0.45 for the BA treatment group and 0.395 for the ST control group, resulting in an odds ratio of 1.14, which is significantly different from 1.00 ($p < 0.05$). In other words, a significantly greater number of patients in the BA treatment group had a change in visual ability that exceeded the MCID clinical endpoint than occurred in the ST control group.

After combining the BA and ST groups, we compared the change in depression severity estimated from Rasch analysis of PHQ-9 responses of patients with MCID = 1 to the change in depression severity of patients with MCID = 0. There was no significant effect in the VITAL study of MCID in visual ability on changes in depression severity (t-test, p = 0.22).

5.5.3 Reducing Rehabilitation Demand: Net Gain from Vision Rehabilitation

The VITAL study and other vision rehabilitation outcome studies that employed the AI [42, 43] agree that on average vision rehabilitation results in a moderate to large size effect of intervention (Cohen's effect size in the range of 0.7 and 1.1). However, as described above for the VITAL study, a recent Cochrane review of randomized controlled trials that compared the effectiveness of different levels or components of vision rehabilitation concluded that additional services beyond the initial low vision consultation produce no or very small incremental effects [44]. In other words, based on current practices there appears to be a diminishing return on investment with increasing amounts of rehabilitation. Thus, to be truly patient-centered we not only want to measure improvements in functional ability, but also measure the utility of those improvements to the patient. To demonstrate how this can be done, even though we still have an incomplete model of a continuous latent variable for the utility of reducing rehabilitation demand, we apply the fabricated parameters that we used for the simulation (listed in Fig. 5.11) to VITAL study outcome data obtained with the AI.

Both importance (I_{nj}) and difficulty (D_{nj}) ratings were obtained on AI Goals at PRE and POST intervention in the VITAL study. Goal items were filtered out (no difficulty rating elicited) if the Goal was rated "not important" (*i.e.*, if $I_{nj} = 0$). Using the simulated "as if" model specified in Fig. 5.11, importance and difficulty ratings of each Goal for each patient at PRE and at POST were replaced with their corresponding part worth utilities (numbers created for the simulation in the green margins of Fig. 5.11). Next, as done for the simulation, the marginal utility of each Goal for each patient at PRE and at POST was computed by taking the product of the assigned part worth utilities. Finally, multi-attribute utilities of totally successful rehabilitation (*i.e.*, utility of reducing rehabilitation demand to zero) were estimated for both PRE and POST intervention Goals for each patient using the Minkowski distance with $b = 2$ (re. left panel of Fig. 5.2), an arbitrarily chosen value that results in attenuation of utility growth with increasing numbers of Goals (the number of Goals with non-zero utilities varied across patients from 2 to 40 at PRE [mean = 16 and SD = 7] and from 2 to 44 at POST [mean = 15 and SD = 9]).

In this hybrid simulation, there is no significant difference in rehabilitation demand (multi-attribute utility) estimated between the BA and ST groups at PRE ($p = 0.344$) or at POST ($p = 0.405$). However, this "as if" model does result in a significant reduction in rehabilitation demand from PRE to POST for both groups

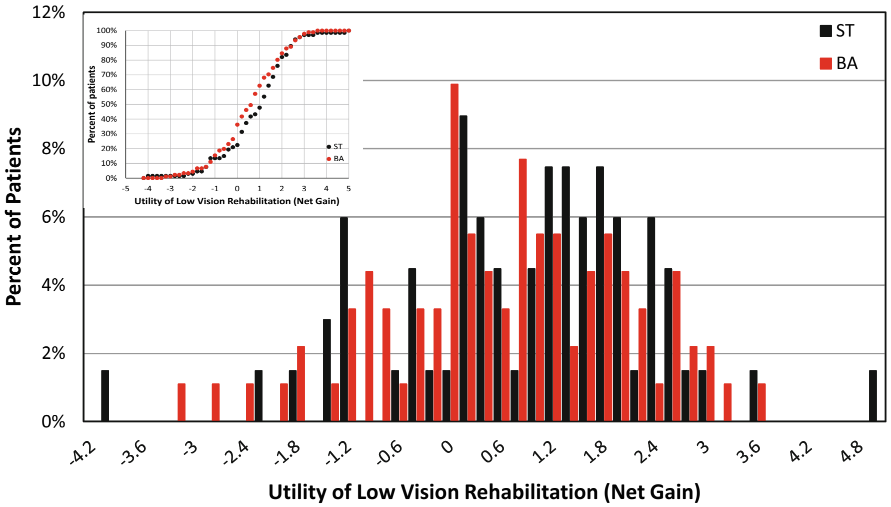


Fig. 5.18 Histogram of the distribution of outcomes of intervention in the VITAL study when the outcome measure is estimated as the utility of rehabilitation demand reduction, estimated with the hybrid (simulation and data) model, from intervention for the ST (black bars) and BA (red bars) groups. Both groups exhibited a significant increase in the utility of visual ability outcomes (decrease in rehabilitation demand). As shown by the cumulative distribution of the outcome measures in the inset, the ST group (control intervention) had slightly better outcomes than did the BA group (experimental intervention), but that difference is not statistically significant

($p = 1.27 \times 10^{-5}$ for ST and $p = 0.00055$ for BA). Figure 5.18 displays histograms of net gains in the utility of vision rehabilitation outcomes (*i.e.*, reductions in rehabilitation demand) for the BA (red) and ST (black) groups. The inset in Fig. 5.18 displays the two distributions as relative cumulative frequency functions. These results would be interpreted as an average reduction in rehabilitation demand of 0.53 (SD = 1.42) for the BA group and 0.82 (SD = 1.5) for the ST group. This difference between groups, however, is not statistically significant ($p = 0.11$).

5.5.4 Next Steps in the Development of Preference-Based Patient-Centered Outcome Measures for Vision Rehabilitation

The above estimates from AI Goal importance and difficulty ratings of multi-attribute utilities representing rehabilitation demand are premature. They were presented here as a demonstration of the next aim in the development of patient-centered outcome measures that incorporate patient preferences. To achieve this aim we ultimately must develop a valid method of estimating the importance of each AI Goal on a continuous interval scale for each respondent (ι_{nj}) that incorporates the stochastic error distributions (ϵ_{nj}). We then must collect sufficient triadic comparison data on a large sample representing the low vision patient population to map continuous importance (ι_{nj}) and difficulty (δ_{nj}) latent variables onto marginal utilities and to define the utility function that maps the part worth utilities onto the total utility for the Goal, $\upsilon_{nj} = u_n(\iota_{nj}, \delta_{nj})$.

Rasch analysis, or some variant of traditional Rasch analysis in the case of importance ratings, must be used to measure the continuous latent variables (ι_{nj} and δ_{nj}) estimated from ordinal ratings of individual patients (I_{nj} and D_{nj}). It then will be necessary to build a large database for a sample of the target low vision patient population to estimate, validate, and anchor model algorithms and parameters for the part worth utility, marginal utility for each Goal, and multi-attribute (rehabilitation demand) utility functions. This theory-driven approach also can give us the tools to identify, estimate, and ultimately understand stochastic and systematic deviations of individual patients from the trends for the targeted population.

A theory-driven approach to the development of patient-centered outcome measurements also promises to provide the tools needed for principled cost-benefit analyses of specific interventions. The ultimate concern to the clinician when assessing risks, costs, and benefits of intervention is the clinical outcome, including adverse events, at a physiological (*e.g.*, ocular pathology) and/or behavioral (*e.g.*, visual impairment) level. The ultimate concern to the patient when assessing risks and benefits of the same intervention is net gains and losses in her or his quality of life, a multi-dimensional construct that ultimately is quantified as a personal multi-attribute utility of the intervention. To facilitate communication between the patient and clinician and thereby facilitate meaningful and ethical shared decision-making, it is necessary to model the relationships between manifest variables observed by the clinician (*e.g.*, visual impairment measures) and latent variables observed by the patient (*e.g.*, visual ability), which we attempt to do with the conceptual (and preliminary computational) model schematized in Fig. 5.7. The clinician has a myriad of sophisticated tools to make objective measurements of publicly observable variables. Rasch models provide us with the tools needed to make objective measurements of latent variables that are observed privately by the patient. What we need now is a rigorous *psychophysics* to build a crosswalk between the two worlds of measurement by way of testable theories.

References

1. R.W. Massof, The measurement of vision disability. Optom. Vis. Sci. **79**, 516–552 (2002)
2. T.G. Bond, C.M. Fox, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 3rd edn. (Routledge, New York, 2015)
3. R.W. Massof, G.S. Rubin, Visual function assessment questionnaires. Surv. Ophthalmol. **45**, 531–548 (2001)
4. R.L. Kirby, The nature of disability and handicap, in *Medical Rehablitation*, ed. by J. V. Basmajian, R. L. Kirby, (Williams & Wilkins, Baltimore, 1984), pp. 14–18
5. C. Bradley, R.W. Massof, Method of successive dichotomizations: An improved method for estimating measures of latent variables from rating scale data. PLoS One **13**, e0206106 (2018)

6. R.W. Massof, Is the partial credit model a Rasch model? J. Appl. Meas. **13**, 1–18 (2012)
7. P.G.W. Jansen, E.E. Roskam, Latent trait models and dichotomization of graded responses. Psychometrika **51**, 69–91 (1986)
8. R.W. Massof, 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 Epidemiol. **18**, 1–19 (2011)
9. G.N. Masters, A Rasch model for partial credit scoring. Psychometrika **47**, 149–174 (1982)
10. F. Samejima, Estimation of latent ability using a response pattern of graded scores. Psychometrika **Suppl 17** (1969)
11. D. Andrich, Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. Psychometrika (2010)
12. G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests* (Danish Institute for Educational Research, Copenhagen, 1960)
13. W.S. Torgerson, *Theory and Methods of Scaling* (Wiley, New York, 1958)
14. E. Muraki, Fitting a polytomous item response model to Likert-type data. Appl. Psychol. Measur. **14**, 59–71 (1990)
15. E.E. Roskam, P.G.W. Jansen, Conditions for Rasch-dichotomizability of the unidimensional polytomous Rasch model. Psychometrika **54**, 317–333 (1989)
16. T. Chan, D.S. Friedman, C. Bradley, R. Massof, Updated estimates of incidence and prevalence of visual impairment, low vision, and blindness in the U.S. JAMA Ophthalmol. **136**, 12–19 (2018)
17. J.E. Goldstein, R.W. Massof, J.T. Deremeik, S. Braudway, M.L. Jackson, K.B. Kehler, S.A. Primo, J.S. Sunness, LOVRNET Study Group, Baseline traits of low vision patients served by private outpatient clinical centers in the United States. Arch. Ophthalmol. **130**, 1028–1037 (2012)
18. R.W. Massof, G. Dagnelie, J.T. Deremeik, J.L. DeRose, S.S. Alibhai, N.M. Glasner, Low vision rehabilitation in the U.S. health care system. J. Vis. Rehabil. **9**, 3–31 (1995)
19. R.W. Massof, A systems model for low vision rehabilitation. I. Basic concepts. Optom. Vis. Sci. **72**, 725–736 (1995)
20. R.W. Massof, A systems model for low vision rehabilitation. II. Measurement of vision disabilities. Optom. Vis. Sci. **75**, 949–973 (1998)
21. J.A. Stelmack, T.R. Stelmack, R.W. Massof, Measuring low vision rehabilitation outcomes with the NEI VFQ-25. Invest. Ophthalmol. Vis. Sci. **43**, 2859–2868 (2002)
22. M. Gobeille, C. Bradley, J.E. Goldstein, R.W. Massof, Calibration of the Activity Inventory item bank: A patient-reported outcome measurement instrument for low vision rehabilitation. Trans. Vis. Sci. Technol. **10**, 12 (2021)
23. R.W. Massof, C.T. Hsu, F.H. Baker, G.D. Barnett, W.L. Park, J.T. Deremeik, C. Rainey, C. Epstein, Visual disability variables. II: The difficulty of tasks for a sample of low-vision patients. Arch. Phys. Med. Rehabil. **86**, 954–967 (2005)
24. E. Wilson, M. Hilferty, The distribution of chi-square. Proc. Natl. Acad. Sci. U. S. A. **17**(12), 684–688 (1931)
25. R.W. Massof, L. Ahmadian, L.L. Grover, J.T. Deremeik, J.E. Goldstein, C. Rainey, C. Epstein, G.D. Barnett, The Activity Inventory: An adaptive visual function questionnaire. Optom. Vis. Sci. **84**, 763–774 (2007)
26. P.G. Moschopoulos, W.B. Canada, The distribution function of a linear combination of chi-squares. Comput. Math. Appl. **10**, 383–386 (1984)
27. R.W. Massof, A clinically meaningful theory of outcome measures in rehabilitation medicine. J. Appl. Meas. **11**, 253–270 (2010)
28. M.A. Goodale, A.D. Milner, Separate visual pathways for perception and action. Trends Neurosci. **15**, 20–25 (1992)
29. G.E. Legge, G.S. Rubin, D.G. Pelli, M.M. Schleske, Psychophysics of reading – II. Low vision. Vis. Res. **25**, 253–265 (1985)
30. R.W. Massof, An interval-scaled scoring algorithm for visual function questionnaires. Optom. Vis. Sci. **84**, 690–705 (2007)

31. J.A. Stelmack, X.C. Tang, D.J. Reda, D. Moran, S. Rinne, R.M. Mancil, R. Cummings, G. Mancil, K. Stroupe, N. Ellis, R.W. Massof, The Veterans Affairs Low Vision Intervention Trial (LOVIT): Design and methodology. Clin. Trials **4**, 650–660 (2007)
32. J.K. Cooper, T. Kohlmann, J.A. Michael, S.C. Haffer, M. Stevic, Health outcomes: New quality measure for Medicare. Int. J. Qual. Health Care **13**, 9–16 (2001)
33. J.J. Manly, N. Schupf, Y. Stern, A.M. Brickman, M.X. Tang, T. Mayeux, Telephone-based identification of mild cognitive impairment and dementia in a multicultural cohort. Arch. Neurol. **68**, 607–614 (2011)
34. J.M. Lyness, T.K. Noel, C. Cox, D.A. King, Y. Conwell, E.D. Caine, Screening for depression in elderly primary care patients: A comparison of the Center for Epidemiologic Studies-Depression Scale and the Geriatric Depression Scale. Arch. Intern. Med. **157**, 449–454 (1997)
35. A.D. Deemer, R.W. Massof, B.W. Rovner, R.J. Casten, C.V. Piersol, Functional outcomes of the low vision depression trial in age-related macular degeneration. Invest. Ophthalmol. Vis. Sci. **58**, 1514–1520 (2017)
36. J.E. Goldstein, M.W. Chun, D.C. Fletcher, J.T. Deremeik, R.W. Massof, Low Vision Research Network Study Group, Visual ability of patients seeking outpatient low vision services in the United States. JAMA Ophthalmol. **132**, 1169–1177 (2014)
37. R.W. Massof, L. Ahmadian, What do different visual function questionnaires measure? Ophthalmic Epidemiol. **14**, 198–204 (2007)
38. R.W. Massof, A general theoretical framework for interpreting patient-reported outcomes estimated from ordinally scaled item responses. Stat. Methods Med. Res. **23**, 409–429 (2014)
39. R.W. Massof, C.T. Hsu, F.H. Baker, G.D. Barnett, W.L. Park, J.T. Deremeik, C. Rainey, C. Epstein, Visual disability variables. I. The importance and difficulty of activity goals for a sample of low vision patients. Arch. Phys. Med. Rehabil. **86**, 946–953 (2005)
40. T.F. Cox, M.A.A. Cox, *Multidimensional Scaling*, 2nd edn. (CRC Press LLC, Boca Raton, 2001)
41. B.W. Rovner, R.J. Casten, M.T. Hegel, R.W. Massof, B.E. Leiby, A.C. Ho, W.S. Tasman, Low vision depression prevention trial in age-related macular degeneration: A randomized clinical trial. Ophthalmology **121**, 2204–2211 (2014)
42. J. Goldstein, M. Jackson, S. Fox, J.T. Deremeik, R.W. Massof, Low Vision Research Network Study Group, Clinically meaningful rehabilitation outcomes of low vision patients served by outpatient clinical centers. JAMA Ophthalmol. **133**(7), E1–E8 (2015)
43. M. Gobeille, A. Malkin, R. Jamara, N.C. Ross, Clinical outcomes of low vision rehabilitation delivered by a mobile clinic. Ophthalmic Physiol. Opt. **38**(2), 193–202 (2018)
44. R.M.A. Van Nispen, G. Virgili, M. Hoeben, M. Langelaan, J. Klevering, J.E.E. Keunen, G.H.M.B. van Rens, Low vision rehabilitation for better quality of life in visually impaired adults (review). Cochrane Database Syst. Rev. **Issue1**, CD006543 (2020). https://doi.org/10.1002/14651858

Chapter 6
Functional Binocular Vision: Toward a Person-Centered Metric

Maureen Powers and William P. Fisher Jr.

Abstract A research program investigating correctable issues in Functional Binocular Vision (FBV) related optometric variables to responses from a symptom survey and reading test results. The study was mounted with no explicit attention to measurement modeling. Data from this research program were retrospectively analyzed with the aims of evaluating the potential for learning from the existing observations, and for improving the study design in future iterations. Results suggest that the physical and psychological measurements of vision combine into a model of FBV that could be standardized and deployed for use in diagnosing significant numbers of untreated vision problems negatively impacting learning outcomes.

Keywords Binocular vision · Psychophysics · Modeling methods · Standards development

6.1 Vision and Reading

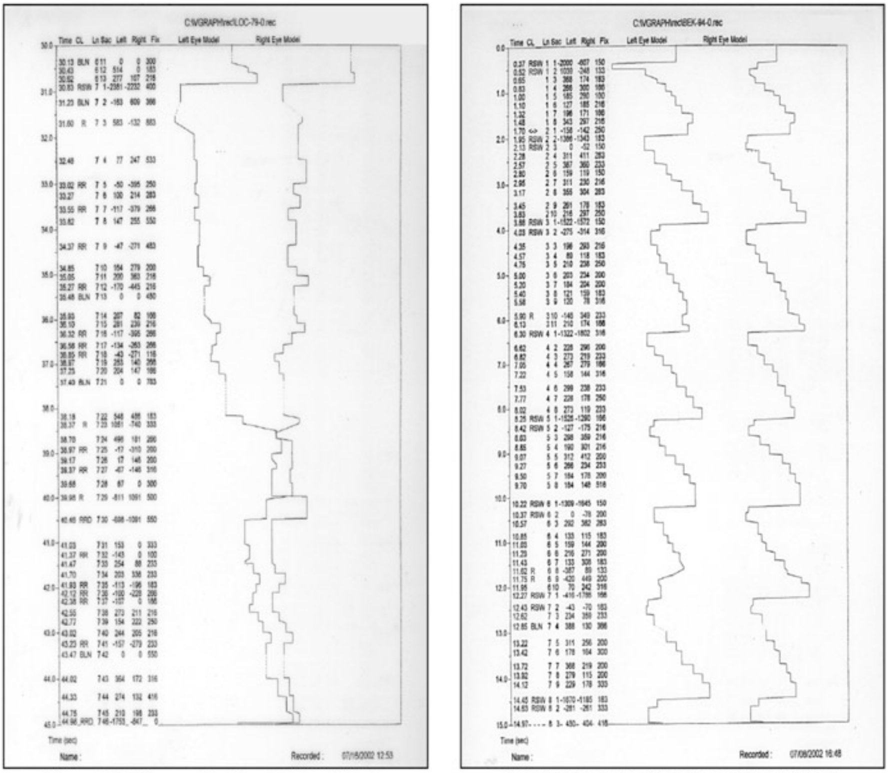
When people think about vision, especially in children, they tend to consider glasses or contact lenses as a first solution – images seem blurry, so better focus should help. Yet there are many reasons vision might not be optimal besides poor focus. Disease, accidents, and genetic factors can all affect the eyes and brain.

Binocularity is one particular aspect of how we process visual information – we have two eyes, each of which presents a slightly different image to the visual cortex, where perception begins. The eyes only "see" because they are connected to the brain. If the two eyes are not well coordinated – if the person's ability to converge or diverge them in order to take in images from different depths in the world, for example, then perception can be inaccurate or misleading.

M. Powers (✉) · W. P. Fisher Jr.
Gemstone Foundation, Rodeo, CA, USA

University of California, Berkeley, CA, USA
e-mail: maureenpowers@gemstonefoundation.org

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3_6



Visagraph of a 14-year-old reading at a 3rd grade level

Visagraph of a 14-year-old reading at a 12th grade level

Fig. 6.1 Visagraph [8] traces of two students' eye movements while reading

Figure 6.1 shows eye movement recordings [8] from two students with very different eye coordination skills. The lines of text being read are on the left; time goes downward. The left trace is from a 14-year-old student (GAR39 59) who exhibits poor vergence control, indicated by the lack of parallel tracking with constant vergence – the eyes drift together and apart. This student's reading efficiency is that of a 3rd grader, even though she is in 8th grade.

The trace on the right is from a different 14-year-old student (GAR28 48) in 8th grade whose reading efficiency is at 12th grade level. The recordings on the right illustrate excellent control, in that both eyes are moving across a line of text from left to right sides of the passage, and then resetting to the beginning of the next line in one smooth movement. The traces are parallel – meaning that the eyes are maintaining a relatively constant convergence and are well coordinated in their movements.

The trace on the left in Fig. 6.1 shows a student who reads 5 grade levels below norms for her grade level because of potentially correctable vision problems, not because of cognitive difficulties. The traced eye movements show they are not coordinated; there are large changes in convergence, which means that the two eyes are not seeing two matched images. At one point they even cross over each other.

One outcome of the kind of poor eye coordination shown on the left in Fig. 6.1 is blurring. When children complain of blurred images even after obtaining corrective lenses, the kind of binocular weakness as demonstrated here can be a cause. Powers and colleagues have shown, using quantitative optometric measures of vergence, accommodation (focusing with the two eyes together), and saccadic tracking [10, 21], that poor visual skills are associated with poor reading outcomes as measured in school-administered tests [22].

6.2 Diagnosing Binocular Vision Problems

In 2003, Borsting and colleagues [5] reported the development of a 15-item Convergence Insufficiency Symptom Survey (the "CISS") intended for children between 9 and 18 years of age (Table 6.1). The items describing the symptoms are read to the subject, who indicates the frequency of occurrence on a scale of 0–4, with 0 representing "Never" and 4 being "Always." Total scores can range from 0 to 60. Borsting et al. [5] established a statistical cutoff from their data, with a score of 16 or higher representing children who exhibited signs of convergence problems from optometric measurements. Those with a score of 15 or lower did not.

Table 6.1 Convergence insufficiency symptom survey item content

Item #	CISS items
1	Do your eyes feel tired when reading or doing close work?
2	Do your eyes feel uncomfortable when reading or doing close work?
3	Do you have headaches when reading or doing close work?
4	Do you feel sleepy when reading or doing close work?
5	Do you lose concentration when reading or doing close work?
6	Do you have trouble remembering what you have read?
7	Do you have double vision when reading or doing close work?
8	Do you see the words move, jump, swim or appear to float on the page when reading or doing close work?
9	Do you feel like you read slowly?
10	Do your eyes ever hurt when reading or doing close work?
11	Do your eyes ever feel sore when reading or doing close work?
12	Do you feel a 'pulling' feeling around your eyes when reading or doing close work?
13	Do you notice the words blurring or coming in and out of focus when reading or doing close work?
14	Do you lose your place while reading or doing close work?
15	Do you have to re-read the same line of words when reading?

Borsting's group later repeated their study with the CISS instrument on adults, and found a higher score was more appropriate to distinguish those with and without convergence insufficiency [24]. Since that time, several groups have found that the survey can also identify accommodation issues (e.g., [15]) which relate to binocularity in that the eyes need to focus on the plane of the object being viewed to see it clearly, via accommodative movements of the intra-ocular lens.

In the field of outcome measurement, professional concern for patient satisfaction has given way to devising means to improve patient involvement [4]. Analysis of patient satisfaction surveys revealed that those who are more involved with their care have better outcomes. Thus, patient participation in and engagement with their care have become a focus of efforts in research and practice. In the context of our work, the student is the "patient," so involvement of family member(s) becomes essential [6]. This concern will be taken up in subsequent research building out deepening engagement in relationships along the continuum from informing to consulting to involving to collaborating to empowering [7, 11, 13]. The state of practice concerning binocular deficiencies remains at the level of the need to ensure students and families are informed about their effects on learning outcomes. General public awareness of children's difficulties with eye coordination remains low. Much more must be accomplished to boost knowledge and awareness before any true involvement can be designed or expected.

During a study on the relation between visual skills such as vergence, accommodation, and tracking ability and reading outcomes in Los Angeles Unified School District (LAUSD) the first author obtained a large dataset with symptom scores and optometric values on the same individual students in the age range specified by the CISS; some were longitudinal, before, during, and after an intervention designed to improve binocular function. The second author suggested this might be a rich dataset for beginning a search for a unified Functional Binocular Vision (FBV) variable – a potentially one-dimensional construct around which a more inclusive and comprehensive battery of vision, survey, and assessment instruments might be designed, with the aim of providing practitioners a better indication of how binocular vision issues affect their students or patients.

6.3 Defining Functional Binocular Vision (FBV)

The optometric variables numbered about 30, all recorded at school during specified testing sessions, after permission from parents was obtained. Examples of measurements taken are:

1. visual acuity measured in each eye separately with the Snellen eye chart placed 20 feet away from the student,
2. the ability to rapidly re-focus the eyes from distances across the room to a book,
3. near point of convergence,
4. vergence ranges and ability to change vergence,

5. tracking ability, and
6. resting eye alignment.

For purposes of analysis, there were 4 items concerning basic optometrics (measuring acuity and ocular balance; both static variables) 7 items addressing vergence (measuring different attributes of dynamic convergence and divergence ability), 4 items for accommodation (different attributes of focusing ability, also a dynamic variable), and the CISS's 15 items concerning symptoms (measuring discomfort and visual symptoms while reading). Clinicians currently select a subset of these variables to determine whether a patient has binocular vision problems, doing so in the absence of any substantive model of the multidimensional phenomenon. We wanted to simplify that task, estimate the model parameters, and perhaps improve diagnostic accuracy, for clinicians and also for testing in schools. Frequencies of the optometric measures were used to assign arrays of ratings in accord with the clinical inferences typically made concerning FBV. Data were fit to a probabilistic model formulated separately for each of the two visual dimensions and one survey dimension measured.

6.4 Methods

After administering the CISS in four elementary schools in LAUSD to 1062 students ages 8 through 11, those with scores greater than 15 were identified for inclusion in the training vision skills study, as this is the criterion indicating convergence problems. Sample size was thus reduced to 418. An FBV scale of 13 items and an acuity scale of eight items were then organized from optometric variables measured for 312 cases overall. The 312 students' repeated training sessions produced a grand total of 4064 individual measurements. Because students varied in the numbers of training sessions they experienced, data from the earliest available and latest available measurements were grouped for the comparisons shown here. In the end, statistical comparisons focused on 76 students with visual skill problems upon assessment who received visual skills training via computer over a period of several months.

Training sessions offered five computerized modules (listed in Table 6.2); the number administered to individual students ranged from one to five across as many

Table 6.2 Module descriptions

Number	Module name	Module purpose
1	Fast focusing	Improve accommodative facility
2	Smooth tracking	Improve smooth pursuit speed and accuracy
3	Jump tracking	Improve saccadic eye movement speed and accuracy
4	Cross-Eyed fusion	Improve convergence range
5	Wall-Eyed fusion	Improve divergence range

Table 6.3 Items by module

Number	Code	Focus	M1	M2	M3	M4	M5
1	AHT2	Average Hit Time	X	X	X	X	X
2	BHT2	Best Hit Time	X	X	X	X	X
3	OHP2	Overall Hit Percent	X	X	X	X	X
4	THR2	Tracking Hit Rate		X	X		
5	RTN2	Red Transition Number	X				
6	BTN2	Blue Transition Number	X				
7	RTA2	Red Transition Average	X				
8	BTA2	Blue Transition Average	X				
9	MB2	Maximum Break				X	X
10	MR2	Maximum Recovery				X	X

as 50 sessions, though most had between 10 and 40. Modules were always administered in the same order. Ten total items were tracked across the five modules, with four to seven included in any given module (see Table 6.3). Time data were recorded in seconds, hit percentages as fractions to four decimal places, and as other kinds of computed scores occurring in various numeric ranges. Applying methods also used in the measurement of chronic disease conditions in clinical chemistry [9], each distribution was divided into eight ranges and 1–8 ratings were assigned, where 1 indicated the worst possible performance, and 8, the best. One item's highest range was between 0.99990 and 1.00000; because scoring was limited to four decimal places and no distinctions could be made, these were scored 7.

Left and right visual acuity optometry measurements were each scored in four categories, with worst at 1 and best at 4. Another six dichotomous items indicated whether or not the student had ever had eye surgery, had ever been diagnosed as needing prescription lens glasses, owned prescription lens glasses, had them with them, wore them in class, was wearing them at this moment.

Individual response data from the reading assessments were not available, so reading scores were incorporated in comparisons as received, with no scaling model applied.

6.5 Measurement Models

The FBV, Acuity, and visual symptoms constructs were each modeled as independent structurally invariant dimensions [23, 27, 28]). The physical and psychological constructs are conceptualized within a common mathematical frame of reference as being measured in interval units offering the potential for metrologically traceable quality assurance standards [14, 16, 17]. The FBV, Acuity, and CISS instruments were then scaled separately, applying rating scale models [2, 3, 29] using the Winsteps software [12], and then were combined in a regression model to predict reading scores.

6.6 Overall Scaling Results

Table 6.4 shows the summary statistics from the separate scaling analyses of the ten-item FBV, eight-item acuity, and 15-item CISS scales for the total samples from which the 312 cases were drawn. The 76 cases in the contrasting Effective and Ineffective groups were then selected from that subset of 312.

Table 6.5 shows the summary item statistics for each scale. The original five categories for CISS ratings were reduced to three by combining Infrequently with Sometimes, and Often with Always, in order to align higher ratings with higher measurements. After this rescoring, measurements were associated with the expected Never, Infrequently/Sometimes, and Often/Always categories for 75%, 56%, and 67% of the responses, respectively, and the observed categories were associated with the expected measurement ranges for 63%, 77%, and 25%, respectively.

Figure 6.2 gives the sample sizes for training sessions completed by our subjects. Note that each module is reported separately. Modules addressed: (1) accommodation facility, (2) smooth tracking, (3) saccadic tracking, (4) convergence, and (5) divergence. The numbers of computerized training sessions were aggregated into 1–7, 8–12, 13–19, and 20 or more for the purposes of later regression analysis. The number of measures summarized ranged from about 150 to 260, with about 750 to 1250 measures within a session group.

Table 6.4 FBV, Acuity, and CISS scale student measurement summary statistics

Scale	# Measurements	Mean/SD responses	Mean/Adj SD[a] measurements	Mean/SD outfit Mn Sq	Measurement separation/ Reliability
FBV	4064	5/1.1	0.08/0.40	0.99/0.77	1.4/0.67
Acuity	637	7.7/0.9	3.33/2.77	0.71/1.40	1.8/0.76
CISS	1412	14.9/0.6	−1.02/1.55	1.00/0.47	2.3/0.84

[a]The standard deviation is adjusted by subtracting the mean square error from the total variance; the square root of this adjusted variance is the adjusted standard deviation

Table 6.5 FBV, Acuity, and CISS scale calibration summary statistics

Scale	# Items	Mean/SD responses	Mean/Adj SD[a] calibrations	Mean/SD outfit Mn Sq	Calibration separation/ Reliability
FBV	10	2038/1350	0.00/0.14	0.97/0.31	9.2/0.99
Acuity	8	611/20	0.00/1.82	0.78/0.52	9.9/0.99
CISS	15	1400/5	0.00/0.56	1.00/0.14	11.2/0.99

[a]The standard deviation is adjusted by subtracting the mean square error from the total variance; the square root of this adjusted variance is the adjusted standard deviation

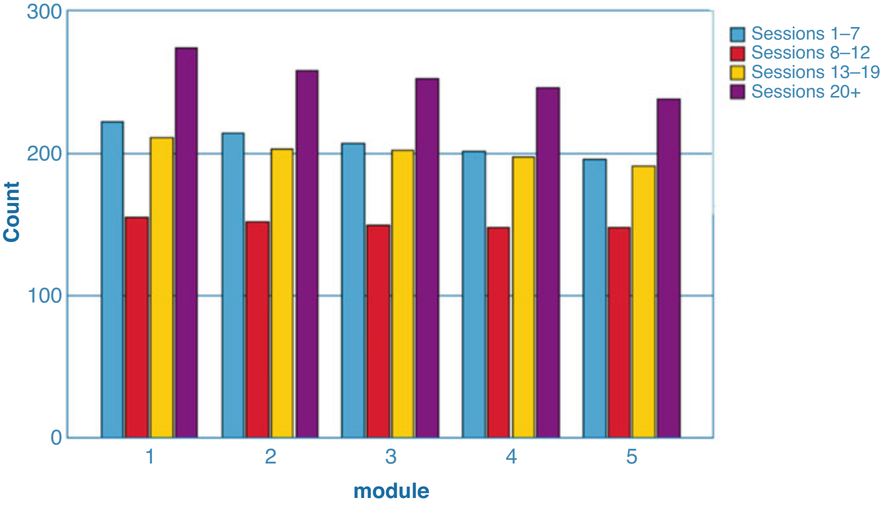


Fig. 6.2 Sample sizes by number of sessions per module

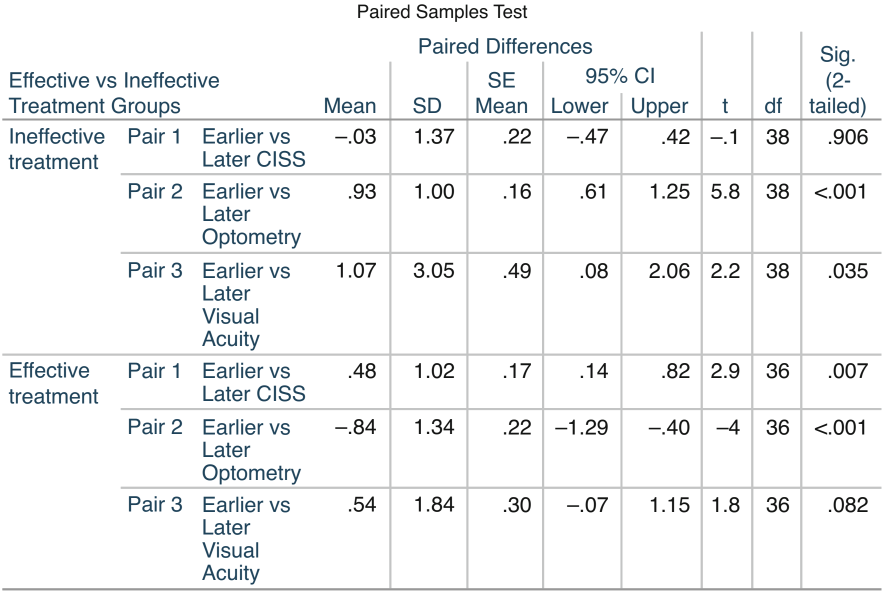
6.7 Changes in FBV with Intervention

The purpose of the study Powers [10, 18, 21, 22] conducted in the Los Angeles schools was to see whether reading fluency scores would improve with systematic intervention by a visual skills training procedure in class. Thus, the data set also contained multiple measurements of all variables repeated several times during the intervention. Though the number of students for whom data were available was limited (N = 76), the trends are encouraging [19, 20].

For about half of the students, training was defined as “Effective,” meaning that they had become proficient in at least four of the five visual skill training modules offered (n = 37). For the remainder (n = 39), training was “Ineffective,” meaning that they did not attain proficiency.

Figure 6.3 shows paired t-test results for three measures: symptoms (earlier to later CISS scores), scaled optometric variables (items like vergence, accommodation, and tracking) and visual acuity (from eye chart measurements) after the third measurement, comparing the Effective treatment vs Ineffective treatment groups. Results for Ineffective Training (top three rows in Fig. 6.3), where students did not complete training or did not attain a level of competency, demonstrate no significant improvements; in fact, both optometric values and acuity got worse. In contrast, the results for Effective Treatment (bottom three rows in Fig. 6.3), categorized by completion of the program with high levels of skill achieved, did show improvement. Training improved both symptom scores and optometric scores, while not affecting acuity.

We were interested in how FBV-related variables changed over time as well as any changes in acuity. Eyechart acuity was not expected to change with an

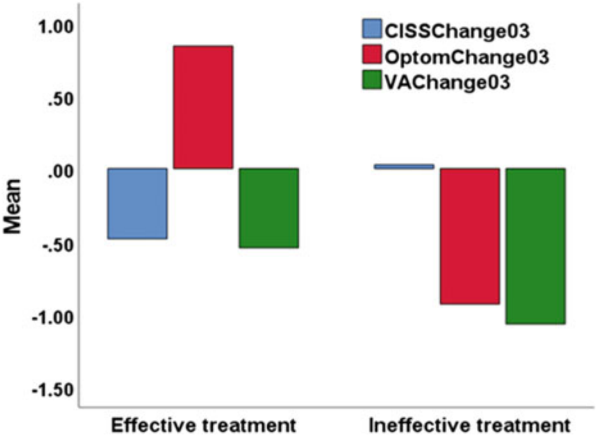


Paired Samples Test

Effective vs Ineffective Treatment Groups			Paired Differences					t	df	Sig. (2-tailed)
					SE	95% CI				
			Mean	SD	Mean	Lower	Upper			
Ineffective treatment	Pair 1	Earlier vs Later CISS	–.03	1.37	.22	–.47	.42	–.1	38	.906
	Pair 2	Earlier vs Later Optometry	.93	1.00	.16	.61	1.25	5.8	38	<.001
	Pair 3	Earlier vs Later Visual Acuity	1.07	3.05	.49	.08	2.06	2.2	38	.035
Effective treatment	Pair 1	Earlier vs Later CISS	.48	1.02	.17	.14	.82	2.9	36	.007
	Pair 2	Earlier vs Later Optometry	–.84	1.34	.22	–1.29	–.40	–4	36	<.001
	Pair 3	Earlier vs Later Visual Acuity	.54	1.84	.30	–.07	1.15	1.8	36	.082

Fig. 6.3 Paired t- test results of differences between measures earlier vs. later in visual skill training

Fig. 6.4 Summary results of Fig. 6.3, showing marked improvement in symptoms (fewer after Effective training) and optometric values (better after training)



intervention that did not attempt to address acuity. The computer program only presented easily detectable targets, well above acuity thresholds, and emphasized the ability to use visual skills like vergence and accommodation- not acuity.

Figure 6.4 is a graph of the data in Fig. 6.3, showing how the mean changed with training for Effective and Ineffective groups. It shows that CISS measurements – where lower is better – declined after intervention. The optometric variables related to binocularity improved after intervention, and visual acuity showed no change- and perhaps worsened over the time period reported.

We concluded at this point that:

- FBV is a uniform variable construct.
- Interventions designed to affect FBV do indeed improve the targeted optometric and symptom responses, but not visual acuity, which was not targeted.
- Scaling variables that have been measured accurately so they can be compared on a common scale allows insights into relationship and mechanisms that would otherwise remain obscure.

6.8 Relationship of FBV to Reading in School Children

One of the goals of our collaboration is to produce a measure of binocular function that relates to reading in school children. FBV appears to do so, at least for this dataset.

The eye movement tracings in Fig. 6.1 are similar to the post-training measurements shown in Table 6.6. Comparing the two, we see that student GAR20 48 had a better outcome in the training measure than student GAR39 59. Similarly, the FBV measure was higher for the good eye movement trace. Students with poor traces scored lower on the reading measure and higher on the symptom measure, but did not differ on the acuity measure from students with good traces.

6.8.1 FBV Measurements Predict Reading Outcomes

Figure 6.5 shows the regression of sustained oral reading scores on FBV measures for a subset of the data in the larger study. The result yields an adjusted R^2 of 0.61, which is highly significant. We can thus conclude that the measures selected to represent FBV (symptoms and optometric measures) are significantly related to this reading outcome measure.

Figure 6.6 shows what happened when visual acuity measurements – which did not change significantly with the intervention – were entered into the regression equation along with FBV measures. Although the regression is still significant, the fit is not better: adjusted R^2 is 0.44, with $p = 0.032$. Thus, visual acuity (a measure of how clear the image appears in each eye individually) contributes less to the change in reading than FBV measures.

Table 6.6 Post-training measurements associated with the Fig. 6.1 eye movement recordings

Student	Mean training measurements across 5 modules	FBV	Acuity	CISS	Reading 1b
GAR20 48	0.33	2.5	4.0	−1.3	5.0
GAR39 59	−0.18	−2.0	4.0	1.5	4.3

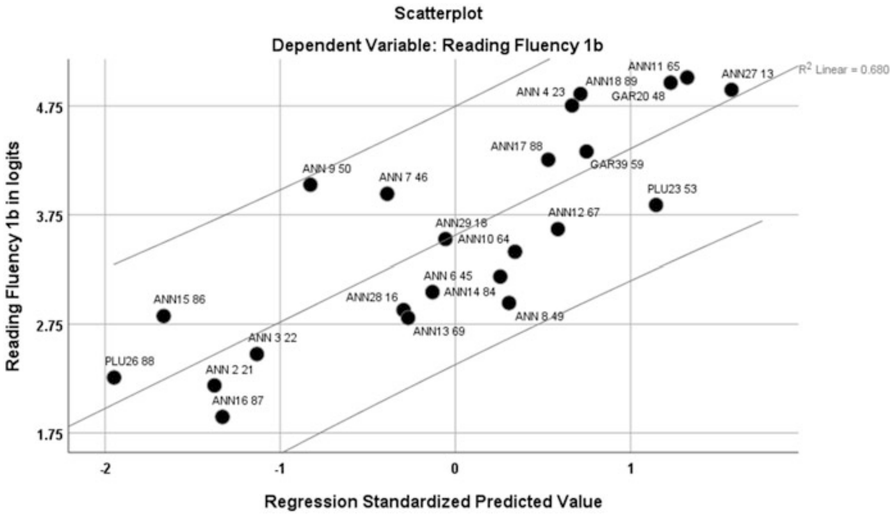


Fig. 6.5 Linear regression of scaled reading fluency (1-min sustained test a grade level) against the FBV symptoms and optometrics variables was significant: R = .825, R^2 = .68, adj R^2 = .61. SE = .6, F change = 9.56, 4 df1, 18 df2, p < .0001. Each point is a subject, N = 23

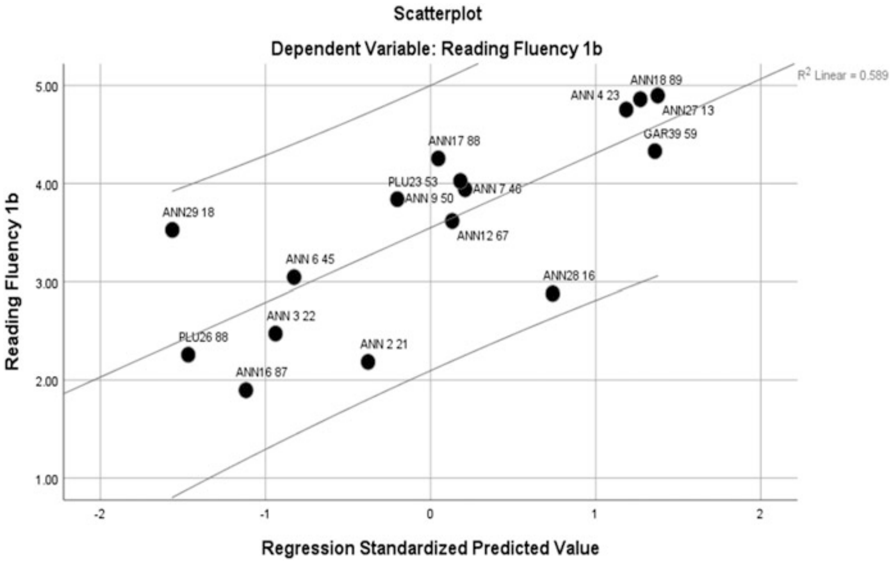


Fig. 6.6 Regression is not as predictive of reading fluency when acuity enters the equation. Plotted are the same FBV variables plus changes in visual acuity. The result is barely significant at the p < .05 level, but the relationship is weaker with acuity included: R = .767, R^2 = .589, adj. R^2 = .44; SE = .74, F change = 3.94, 4 df1, 11 df2, p = .032. N for this comparison is 16 because not all students had complete data

Thus, even though testing visual acuity is important for determining referrals for eye disease or optical needs, it cannot identify problems of eye coordination. This is illustrated in the multidimensional framework sketched in the association of Fig. 6.1 and Table 6.6.

6.9 Limitations

There are some caveats to the interpretation of our results. First among these is that none of the instruments involved were designed with the intention of calibrating interval measurements of FBV. The analyses reported are intended only to provide a basis for learning from the existing data so as to proceed toward improved research designs in the future.

More specifically, the original study was not designed to define FBV in terms of mathematical modeling, so there was no intention to try to gauge the effect of changes in the FBV construct on reading performance. The data were collected during a study designed to see whether training in visual skills such as vergence could improve reading. The individual variables, time points measured, and completeness of each students' data set are limited, therefore, in ways that would not have been allowed in an intentionally designed calibration study.

Also, the N in the regression with acuity is smaller than the N in the regression without acuity. Part of the loss of significance can thus probably be attributed to a loss of power. The fact that acuity did not change when the other measures did, as expected, is probably also relevant.

6.10 Conclusions and Future Directions

1. Functional Binocular Vision (FBV) is a unidimensional construct made up of several measures, including a symptom survey and optometric measurements. It appears to be valid and measurable.
2. More work is needed to determine FBV's reliability and its most meaningful components. However our findings with predicting reading scores suggest that acuity is not related to FBV or on its effects on measures of sustained reading fluency.
3. FBV components are symptoms and optometric measures such as ability to rapidly re-focus the eyes and move them toward and away from the nose in vergence movements. These abilities are reflected in eye movement recordings for good and poor readers, but FBV's demonstrated relationship to fluency suggests other, less invasive measures may help us develop more meaningful testing of vision in our schools.
4. The rating scale models applied in this study can in principle be expanded to incorporate all four constructs in a multidimensional model [1] that leverage the information in the cross-dimensional correlations to improve uncertainty

estimation. We are working toward a user-friendly multidimensional tool for evaluating functional vision. This tool would account for the various aspects of functional binocular vision represented by optometric physics, diagnostic survey data, and reading performance assessments.

This kind of modeling appears to offer a viable approach to establishing person-centered standards for FBV measures, like those that currently exist for acuity. The Snellen [25] eye chart and its derivatives are the gold standard for measuring subjective acuity, which relates to the need for glasses but says nothing as to how well the eyes move. Our goal is to create a model that assesses visual skills with sensitivity and specificity as fit for the purposes of diagnosis and treatment of FBV issues as the Snellen eye chart is for acuity [26]. Such a model and associated measurement standards could be used in conjunction with acuity standards to provide better, more comprehensive, and useful vision testing in children and adults.

Acknowledgments The original data collection for this work was supported by NIH research grant R01 EY017414 to MKP.

References

1. D.D. Allen, M. Wilson, Introducing multidimensional item response modeling in health behavior and health education research. Health Educ. Res. **21**(suppl_1), 73–84 (2006)
2. D. Andrich, A rating formulation for ordered response categories. Psychometrika **43**(4), 561–573 (1978)
3. D. Andrich, Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. Psychometrika **75**(2), 292–308 (2010)
4. P. Bate, G. Robert, Experience-based design: from redesigning the system around the patient to co-designing services with the patient. BMJ Qual. Saf. **15**(5), 307–310 (2006)
5. E.J. Borsting, M.W. Rouse, G.L. Mitchell, M. Scheiman, S.A. Cotter, J. Cooper, et al., Validity and reliability of the revised convergence insufficiency symptom survey in children aged 9 to 18 years. Optom. Vis. Sci. **80**(12), 832–838 (2003)
6. K.L. Carman, P. Dardess, M. Maurer, S. Sofaer, K. Adams, C. Bechtel, J. Sweeney, Patient and family engagement: A framework for understanding the elements and developing interventions and policies. Health Aff. **32**(2), 223–231 (2013)
7. E.M. Castro, T. Van Regenmortel, K. Vanhaecht, W. Sermeus, A. Van Hecke, Patient empowerment, patient participation and patient-centeredness in hospital care: A concept analysis based on a literature review. Patient Educ. Couns. **99**(12), 1923–1939 (2016)
8. D.S. Colby, H.R. Laukkanen, R.L. Yolton, Use of the Taylor Visagraph II system to evaluate eye movements made during reading. J. Am. Optom. Assoc. **69**(1), 22–32 (1998)
9. W.P. Fisher Jr., E. Burton, Embedding measurement within existing computerized data systems: scaling clinical laboratory and medical records heart failure data to predict ICU admission. J. Appl. Meas. **11**(2), 271–287 (2010)
10. D. Grisham, M. Powers, P. Riles, Visual skills of poor readers in high school. Optom. J Am. Optom. Assoc. **78**(10), 542–549 (2007)
11. J. Hibbard, J. Stockard, E. Mahoney, M. Tusler, Development of the Patient Activation Measure (PAM): conceptualizing and measuring activation in patients and consumers. Health Serv. Res. **39**(4, Part I), 1005–1026 (2004)
12. J.M. Linacre, A user's guide to WINSTEPS Rasch-Model computer program, v. 5.1.1. Beaverton, Oregon: Winsteps.com. Retrieved from https://www.winsteps.com/manuals.htm

13. U. Majid, A. Gagliardi, Conceptual frameworks and degrees of patient engagement in the planning and designing of health services: A scoping review of qualitative studies. Patient Exp. J. **6**(3), 82–90 (2019)
14. L. Mari, M. Wilson, A. Maul, *Measurement Across the Sciences* (Springer, Cham, 2021)
15. L.F. Marran, P.N. De Land, A.L. Nguyen, Accommodative insufficiency is the primary source of symptoms in children diagnosed with convergence insufficiency. Optom. Vis. Sci. **83**(5), 281–289 (2006)
16. L.R. Pendrill, *Quality Assured Measurement: Unification Across Social and Physical Sciences* (Springer, Cham, 2019)
17. L. Pendrill, W.P. Fisher Jr., Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. Measurement **71**, 46–55 (2015)
18. M.K. Powers, Improving visual skills: a new internet application. J. Mod. Opt. **53**(9), 1313–1323 (2006)
19. M.K. Powers, W.P. Fisher, Advances in modelling visual symptoms and visual skills. J. Phys. Conf. Ser. **1379**(1), 012044 (2019) IOP Publishing
20. M.K. Powers, W.P. Fisher Jr., Physical and psychological measures quantifying functional binocular vision. Meas. Sens. **18**, 100320 (2021)
21. M. Powers, D. Grisham, P. Riles, Saccadic tracking skills of poor readers in high school. Optom. J. Am. Optom. Assoc. **79**(5), 228–234 (2008)
22. M.K. Powers, J.D. Grisham, J.K. Wurm, W.C. Wurm, Improving visual skills: II—Remote assessment via Internet. Optom. J. Am. Optom. Assoc.**80**(2), 61–69 (2009)
23. G. Rasch, Probabilistic models for some intelligence and attainment tests (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut (1960)
24. M.W. Rouse, E.J. Borsting, G.L. Mitchell, M. Scheiman, S.A. Cotter, J. Cooper, et al., Validity and reliability of the revised convergence insufficiency symptom survey in adults. Ophthalmic Physiol. Opt. **24**(5), 384–390 (2004)
25. H. Snellen, *Probebuchstaben zur Bestimmung der Sehschärfe* (Utrecht, 1862)
26. L. Tong, S.M. Saw, D. Tan, K.S. Chia, W.Y. Chan, A. Carkeet, et al., Sensitivity and specificity of visual acuity screening for refractive errors in school children. Optom. Vis. Sci. **79**(10), 650–657 (2002)
27. B.D. Wright, Solving measurement problems with the Rasch model. J. Educ. Meas. **14**(2), 97–116 (1977)
28. B.D. Wright, A history of social science measurement. Educ. Meas. Issues Pract. **16**(4), 33–45, 52 (1997)
29. B.D. Wright, G.N. Masters, *Rating Scale Analysis* (MESA Press, Chicago, 1982)

Chapter 7
Advancing the Metrological Agenda in the Social Sciences

John Michael Linacre

Abstract For over 100 years social scientists have been measuring their subjects on scales regarded as linear. An early example is “The Thorndike Scale for Handwriting of Children” (Thorndike EL. The Thorndike Scale for Handwriting of Children. Bureau of Publications – Teachers College, Columbia University, New York, (1912)), where it is said that “The unit of the scale equals approximately one-tenth of the difference between the best and worst of the formal writings of 1000 children in Grades 5–8.” Though the construction of linear scales in social science has become more rigorous, an early feature continues. Each measurement scale represents a unique perspective on the target variable. Measurement scales for the same or similar target variables are rarely connected. This chapter describes how measurement units for similar scales can be aligned so that measures become independent of the specifics of the situation on which they are based.

Keywords Estimation methods · Instrument equating · Unit definition · Phobias

7.1 Building a Foundation

The data we are considering here are usually a set of ordinal responses or ratings by a sample of persons to an instrument, which is a set of questions or test items. Typically, the items are carefully written and are appropriate for the targeted people. The data are collected and then statistical manipulation is performed to extract from those data a numerical scale for the intended latent variable. The extraction is considered successful if the statistical model employed is deemed to describe the data satisfactorily.

J. M. Linacre (✉)
The University of the Sunshine Coast, Sunshine Coast, Queensland, Australia
e-mail: mike@winsteps.com

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3_7

As a first step away from the specifics of this type of scaling, we need measurement models which not only surmount the idiosyncrasies of individual datasets, but also intentionally construct linear measures as independent as statistically possible of the specifics of the items and persons generating the data. Then departures in the data from these measurement models become failures in the data to support generalizable linear measurement, rather than failures in the statistical model to describe the characteristics of the data exactly. This is one crucial distinction between statistical models, which describe the data, and measurement models, which prescribe the data, and there are other distinctions [7].

Of course, empirical data always fall short of an ideal, so the criterion for acceptance of data points becomes their utility. For instance, suppose a student makes several lucky guesses on a multiple-choice arithmetic test. Social fairness may mandate that the student be credited with these successes and measured accordingly. However, this measurement may result in the student being placed in an advanced arithmetic class for which the student is not prepared. An analogous situation when measuring a child's height would be that the child stood on tip-toe at the moment of measuring. The child's extra height might allow the child to ride a roller-coaster on which the child cannot be restrained safely. However, if a student must be credited with lucky guesses, an immediate solution is to disregard the lucky guesses while constructing the measurement scale so that guessing does not distort the scale itself. Then, when measuring each student on this scale, lucky guesses are allowed to increase the students' measures, but are flagged and brought to the attention of the student, teacher, parent, etc., utilizing tools like the "Kidmap" [5, 45], so that the responses' status as exceptional and unexpected can inform instruction.

Models which meet our requirement for intentional linear measurement are the Rasch model [30] and its extensions [1, etc.]. These construct linear measures from a correspondence of ordinal data with explanatory theory, and also support the examination of individual data points (responses, ratings, etc.) for their ability to support linear measurement of a latent variable. Once the empirical capacity to obtain data of the needed quality is established, then the latent variable itself can be investigated as to whether it facilitates fit-for-purpose comparisons. However, a form of data-dependency remains in any local scaling application in that each set of linear measures constructed from different instruments and groups of respondents has its own origin (zero-point) and unit size, based on the logit, as these emerge from the given dataset. Logits are a probabilistic unit whose substantive size depends on its context [21]. In applying a model of measurement facilitating the construction of linear measures for each instrument and so quantifying a given measurand in a particular unit, we have advanced, but not far enough.

7.2 The Need for a Universal Measurement Scale for a Variable

"Un roi, une loi, un poids, et une mesure" (One king, one law, one weight, and one measure) was an early slogan of the French Revolution [12]. It summarized the situation that, although all the weights and measures were linear, they differed in different locations and for different classes of people. The French peasants were victims of this system. Consequently, in 1799, the metric system was launched in France to generalize all weights and measures.

Disparities between measurement scales in the social sciences have a less obvious economic and social impact, but we can see their effect [8, 28]. Physical science, based on generalized measures, is advancing rapidly. Findings can be shared easily and productively. On the other hand, social science journals report endlessly on the construction of local measurement scales, but even when the opportunities present themselves, these scales are not combined together to aid the advancement of social science. Thurstone [38, p. 10] understood that mathematical modeling and measurement are not merely useful tools, but should be the very language in which one thinks. We then unnecessarily encumber our communications and ability to learn when we fail to link our measures together in common languages. But even when linear measures are obtained, current practice in psychology and the social sciences reports results from instruments measuring the same thing in different units. The following illustrations are intended to show that an ever-growing Tower of Babel need not be taken for granted. On the contrary, correspondences between different instruments measuring the same thing can be leveraged systematically so as to more explicitly link new results with old.

7.2.1 Our Objective: Combining Local Measurement Scales

Let us demonstrate methods for combining local measurement scales into universal measurement scales using the example of Phobia. According to Kessler et al. [13], 18.1% of adults in the U.S.A. suffer from anxiety disorders. Phobias are a common form of anxiety disorders, so let us construct linear measures of phobia intensity. We will do this using the dataset from Imaizumi and Tanno [11], but will ignore their analysis and findings. The dataset consists of the responses of 582 Japanese adults to 17 questions (symptoms) relating to their experience of Trypophobia. Only one symptom directly mentions this specific phobia, so here it is slightly rewritten to apply to any phobia. The responses were on a rating scale from 1 ("Not at all") to 5 ("Extremely"). Our aim is to construct a measurement "ruler" that can be applied to any phobia by anyone. In so doing, we will show how satisfaction of the invariance, unidimensionality, and construct definition requirements of measurement set the stage not only for equating the instruments to a common metric, but for metrological traceability and quality assurance.

Table 7.1 Seventeen phobia symptoms and their usage in the samples analyzed

	Used in sample				
Symptom number	1	2*	3†	One word	Symptom label
1	X		R	Freaked	Freaked out
2	X		B	Disgust	Aversion, disgust or repulsion
3	X	X	B	Uneasy	Uncomfortable or uneasy
4	X	X	R	Panic	Panicking or screaming
5	X	X	M	Anxious	Anxious, full of dread or fearful
6			M	Nauseous	Sick or nauseous
7			T	Nervous	Nervous
8			M	Crazy	Feel like going crazy
9			T	Urge	Have an urge to destroy the cause
10	X	X	.	Itchiness	Feel itchiness
11		X	B	Skin	Feel skin crawl
12		X	.	Bumps	Have goosebumps
13		X	T	Crying	Feel like crying
14	X	X	T	Vomit	Vomit
15	X		B	Chills	Get chills
16	X		.	Breath	Have trouble breathing
17	X		M	Shiver	Shiver

Note: X = this symptom selected for this Sample; 2* = the 5-category rating scale is dichotomized to two categories; 3† = R, B, M, T are symptom-selection codes, see text

Table 7.1 shows the 17 phobia symptoms in the dataset, together with a one-word abbreviation of each symptom. In this chapter, three separate samples of clients are extracted from the dataset, along with their responses to three different, but overlapping, subsets of questions. The symptoms for each sample are identified in Table 7.1, where "X" or a code letter indicates that this symptom is selected for the sample.

Each sample and subset of questions mimics the collection of data on separate questionnaires with some equivalent questions. The data from each sample will be analyzed with different estimation methods implemented in different software. Based on the estimates (measures) for the three samples, a linear measurement scale of all 17 symptoms will be constructed which could be the basis for combining further phobia questionnaires or constructing new ones. Using this measurement scale, client measures based on any subset of the 17 symptoms can be expressed as measures on the scale of all 17 symptoms, demonstrating that the measures can be independent of the particular subset of questions answered.

This demonstration focuses on the technical aspects of constructing and combining measurement scales. In a full implementation, we would require content experts, such as psychologists, to verify that we are measuring what we intend to measure (Construct Validity) and that the measures make sense when applied to clients (Predictive Validity).

7.2.2 Constructing a Local Measurement Scale for a Latent Variable

Sample 1 is the first 200 clients with non-extreme scores in the Trypophobia dataset. Clients with extreme scores, whose responses are all in the top category of the rating scale or all in the bottom category, are uninformative for constructing a measurement scale because they do not differentiate between the severity of the symptoms.

We will construct linear measures for these 200 clients and the 10 symptoms indicated in Table 7.1. Client phobia intensities and phobia symptom severities are located on the same measurement scale (conjoint measurement) using the probabilistic Rasch Rating Scale Model [1] and Marginal Maximum Likelihood Estimation (MMLE) implemented in the TAM software [33]. Each symptom is modeled to have one "symptom severity" parameter, and each client is modeled to have one "phobia intensity" parameter. There is also a 5-category rating scale from 1, "Not at all", to 5, "Extremely". It is modeled with 4 parameters, the Andrich thresholds, with one for each point of equal probability of adjacent categories.

Our working hypotheses are that the clients with the highest scores on the symptoms have the highest phobia intensity, and that the symptoms with the highest client scores are those most often experienced, and so are the least severe indicators of phobia. These hypotheses will need to be confirmed by phobia experts; the goal here is only to see if data from a professionally constructed measure of phobia exhibit the patterns of invariance necessary and sufficient to estimating quantity values independent of which symptoms are used to measure the clients and which clients are measured.

In MMLE, the client phobia intensity measures are modeled to have a normal distribution, locally centered on zero. Since client measures are never exactly normally distributed, this may slightly bias the measures, but this bias is inconsequential for most practical uses, as we will see. An essential aspect of the methodology presented here is that findings are always provisional. We never know the exact truth, but we do need our findings to be "good enough for government work," which is to say, good enough to accomplish the task at hand to the required tolerance levels as efficiently as possible.

Table 7.2 shows the results of the MMLE estimation. The "Symptom Number" is the original question number in the Trypophobia instrument. The Symptom Measures estimated by TAM are in logits, reported with three decimal places. The zero point is the mean phobia intensity of this sample of clients so that roughly half the clients would be reported with negative phobia intensity.

These logit measures, with negatives and decimals, can be difficult for a non-technical audience to understand and use. We are also intending to generalize this scale to other symptoms, which may be more or less severe than any of these symptoms. New severity extremes would cause the added symptoms to push the logit scale to more negative and/or positive values. This effect might possibly also cause the zero midpoint to be pushed up or down the scale, rendering the numeric values even less intuitive than they originally were.

Table 7.2 Rasch severities of symptoms for Sample 1

Symptom number	Score	Logit calibration	Logit uncertainty	Rescaled calibration	Rescaled uncertainty	One word
1	492	0.799	0.089	466	5	Freaked
2	643	−0.342	0.086	406	4	Disgust
3	624	−0.200	0.086	414	4	Uneasy
4	384	1.751	0.100	515	5	Panic
5	462	1.043	0.091	479	5	Anxious
10	377	1.823	0.102	519	5	Itchiness
14	277	3.159	0.137	589	7	Vomit
15	387	1.721	0.100	514	5	Chills
16	319	2.500	0.116	554	6	Breath
17	309	2.639	0.120	562	6	Shiver

Note: Conversion: Rescaled Calibration = 52.1 * Logit Calibration + 424.2

Accordingly, we will linearly rescale the measurement range for the current symptoms and clients so that all the client phobia measures can be reported as convenient positive integers on a linear scale from 200 to 800. For the symptoms, these rescaled symptom calibrations are in column 5 of Table 7.2. The conversion shown in Table 7.2 is Rescaled Calibration = 52.1 * Logit Calibration + 424.2. Figure 7.1 shows histograms of the clients and symptoms on the 200–800 Scale.

We are delighted to see that many clients in this sample have only milder symptoms and few have more severe symptoms. A measure of 200 corresponds to a client who reports “not at all” to all 10 symptoms. A measure of 800 corresponds to a client who reports “Extreme” to all 10 symptoms.

The rating scale has 5 categories from 1 (not at all) to 5 (extremely). Can our respondents discriminate 5 levels of intensity of their phobias? As we move from left to right in Fig. 7.2, it shows the increasing, then decreasing, probability of observing each category for this sample of clients as inferred by the Rasch model. The key feature in Fig. 7.2 is that each category in turn becomes most probable to be observed. This confirms that the persons in Sample 1 are able to discriminate 5 levels of symptom intensity. Good!

In Fig. 7.2, these four Andrich transition thresholds, which are the locations of equal probable adjacent categories, have the rescaled values of −89, −24, 23, and 90 relative to the severity of the symptom. This means that, when a client’s measure matches a symptom’s calibration, so that there is a 0 difference, a vertical line at 0 in Fig. 7.2 shows that a response in category 3 would occur 40% of the time, responses in categories 2 or 4 would each occur 25% of the time, and responses in categories 1 or 5 would each occur 5% of the time. Similar patterns of probabilities summing to 100% can be read from Fig. 7.2 for instances in which the client providing the ratings has a measure more or less above or below any given symptom.

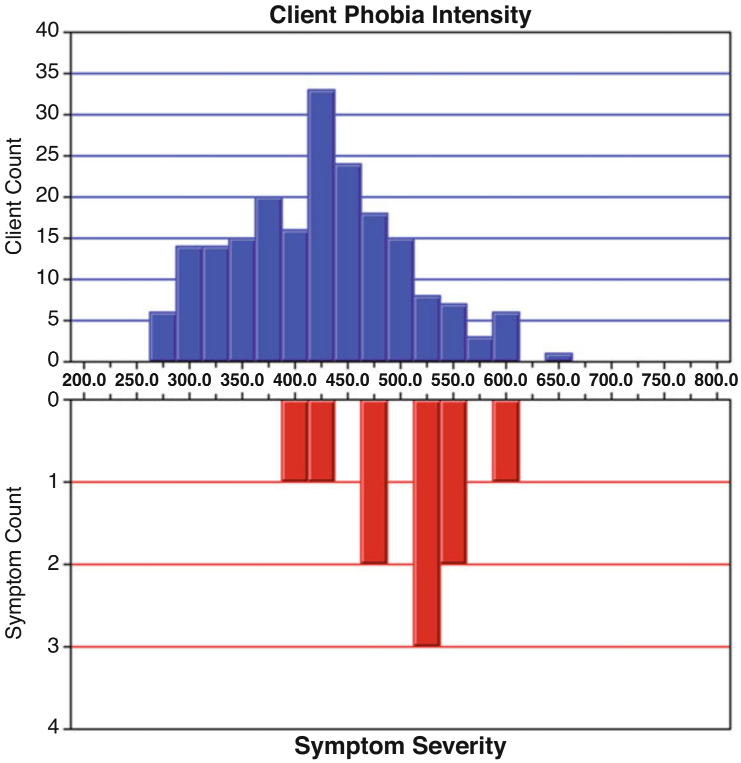


Fig. 7.1 Distributions of the 200 clients in Sample 1 and their 10 symptoms on the 200–800 phobia scale

7.2.3 *Equating Sample 2 of Clients*

The data analyzed for Sample 1 were collected using a specific 5-category item format. Another instrument may use a different format, such as a different rating scale. Let's simulate this using Sample 2, comprised of another 100 clients from the Trypophobia dataset, with 8 of the 17 symptoms (see Table 7.1), four of which were also included in Sample 1.

But instead of retaining the five rating scale categories, we will dichotomize them into two categories to simulate a fictional situation in which another instrument designer approached the phobia construct from their own point of view, with their own purposes. Now, the original categories 1 and 2 are rescored 0, mild. Original categories 3, 4, and 5 become 1, severe. This data will be analyzed in eight ways using a variety of estimation methods and software packages. We will also analyze Sample 2 with the original 5 categories for comparison.

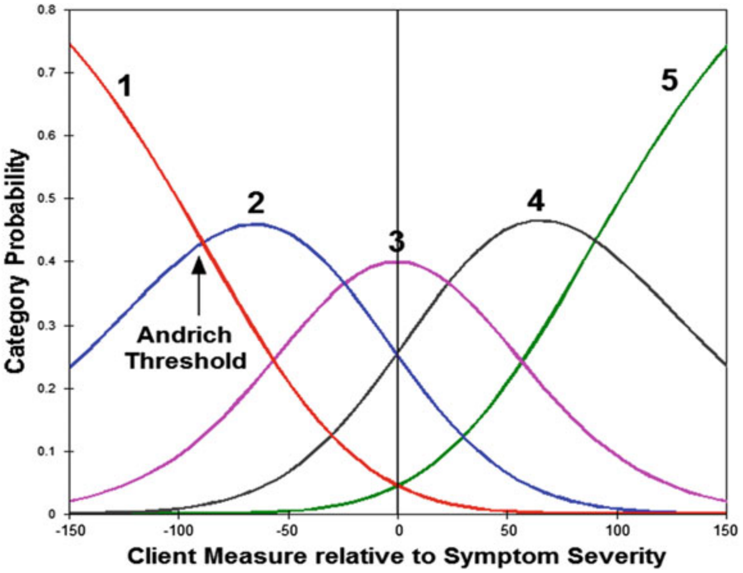


Fig. 7.2 Rasch-model probability curves for the Sample 1 phobia rating scale. Andrich thresholds are at the points of equal probability of adjacent categories. They are −89, −24, 23, and 90, relative to the severity of the symptom. At the vertical line, the client phobia intensity equals the symptom phobia severity

The overall intent is to demonstrate how we can free our thinking from the dictates of results that vary across samples in their specific features but which are actually statistically identical. We focus attention on the patterns of invariance from shared structures capable of supporting instrument equating, metrological traceability, and quality-assured quantity values. Here are the nine estimation methods.

1. Conditional Maximum Likelihood Estimation (CMLE) using "eRm" [27]. CMLE is regarded as the best estimation method because its item difficulty (symptom severity) estimates are statistically consistent [40]. However, CMLE is severely limited in the data designs that it can estimate. For instance, today's datasets are often large and sparse. These are inestimable by CMLE, but can be estimated by other methods such as MMLE and JMLE. CMLE estimates, furthermore, are not symmetric. The distance on the measurement scale between a Symptom and a Client depends on whether Symptoms or Clients are conditioned out of the estimation equations. Here we first condition out the clients, then estimate the symptom calibrations in logits. These logits are reported in Table 7.3.
2. CMLE using "eRm" is then applied to the transposed Sample 2 data matrix to estimate logit measures for the clients (columns) by conditioning out the symptoms (rows). We then anchor (fix the values of) the client measures and use them to estimate the symptom calibrations using Anchored Maximum Likelihood Estimation (AMLE), also called MLE, implemented in "eRm".

Table 7.3 Comparison of logit symptom severity Calibrations for Sample 2

		Logit Calibration									
Symptom No.	Score	CMLE	AMLE	JMLE1	JMLE2	MMLE1	MMLE2	PMLE1	PMLE2	JMLE5	Sample 1 rescaled calibration
3	79	−2.36	−2.04	−2.88	−1.80	−1.60	−1.68	−1.54	−2.12	−1.66	414
4	23	0.83	1.83	0.99	1.61	1.46	1.54	1.17	0.85	0.55	515
5	50	−0.74	−0.03	−0.87	−0.02	0.00	0.00	−0.44	−0.83	−0.48	479
10	28	0.49	1.44	0.60	1.27	1.15	1.21	0.93	0.42	0.37	519
11	54	−0.96	−0.28	−1.12	−0.25	−0.20	−0.21	−0.60	−0.92	−0.92	.
12	65	−1.59	−1.01	−1.85	−0.89	−0.76	−0.80	−1.37	−1.69	−1.11	.
13	6	2.58	3.88	3.05	3.42	3.20	3.34	2.87	2.21	2.00	.
14	12	1.76	2.92	2.08	2.57	2.37	2.48	2.31	2.09	1.25	589

3. JMLE1: Joint Maximum Likelihood Estimation (JMLE) with “Winsteps” [20] is used to estimate both the Symptoms and the Clients simultaneously. JMLE estimates are symmetric, so the extra steps taken in the CMLE examples are not needed.
4. JMLE2: JMLE with “TAM” again estimates both the Symptoms and the Clients simultaneously, using the same method but different software.
5. MMLE1: MMLE with “ltm” [31] is used to estimate the symptom severities. MMLE assumes a normal distribution of clients.
6. MMLE2: MMLE with “TAM” is used to estimate the symptom severities.
7. PMLE1: Pairwise Maximum Likelihood Estimation (PMLE) with “sirt” [32] is used to estimate the symptom severities. PMLE is statistically consistent and allows great flexibility in data designs. However, its uneven use of observations in the data matrix can lead to situations in which estimates are biased.
8. PMLE2: PMLE with “pairwise” [9] is used to estimate the symptom severities.
9. JMLE5: JMLE with “Winsteps” is used to estimate symptom severities for Sample 2 data in its original 5-category format.

These nine sets of logit calibrations are shown in Table 7.3 and are plotted in Fig. 7.3. We can see immediately that the local logit scaling of each analysis produces findings that are promising indications of convergence on a common construct and scale, but which are simultaneously confusing expressions of different

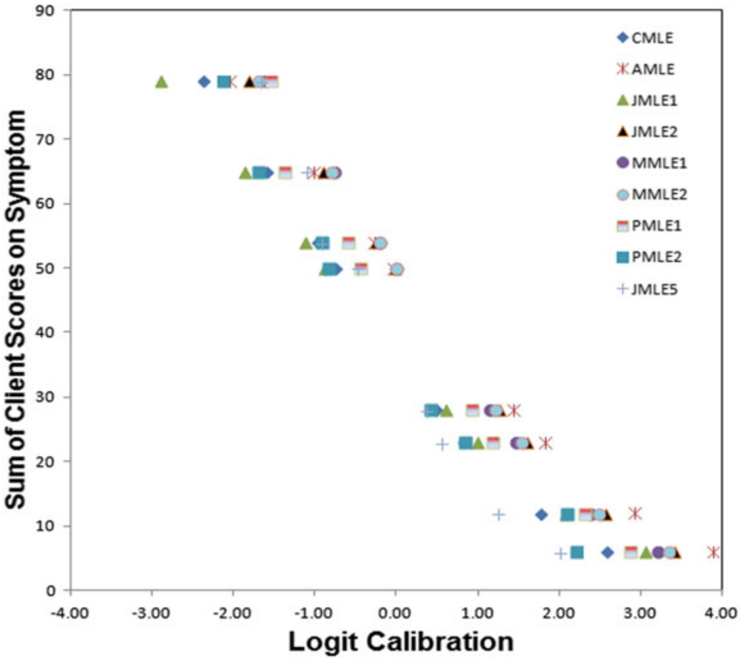


Fig. 7.3 Plot of symptom scores against logit symptom severity calibrations for Sample 2

and not easily compared numeric values. In Fig. 7.3, each symptom is one row located vertically by the symptom's score, which is the same for all estimation methods except the 5-category JMLE5. For convenient comparability, the 5-category JMLE calibration has been placed on the row for the symptom's dichotomous score.

In Fig. 7.3, there is a hierarchy of symptom severities held in common across the estimation methods. Symptoms with lower scores are more severe and so have higher logit calibrations, but the logit ranges for the symptoms estimated by the various methods overlap so that there is no precise picture of the hierarchy of symptoms. This is despite the fact that the "logit" itself has a precise probabilistic definition. That definition, however, results in logits having different substantive expressions depending on the characteristics of the data, the estimation method, and the implementation of that estimation method [21].

To overcome the deficiencies of local logit scaling, all 9 sets of calibrations in Table 7.3 are rescaled onto the 200–800 range defined by Sample 1. The empirical justification for this rescaling is shown in Fig. 7.4, which plots the rescaled calibration for each shared symptom from Table 7.2 against the logit calibrations of Table 7.3. Symptom 4, "Panic", has calibrations that are discordant between Sample

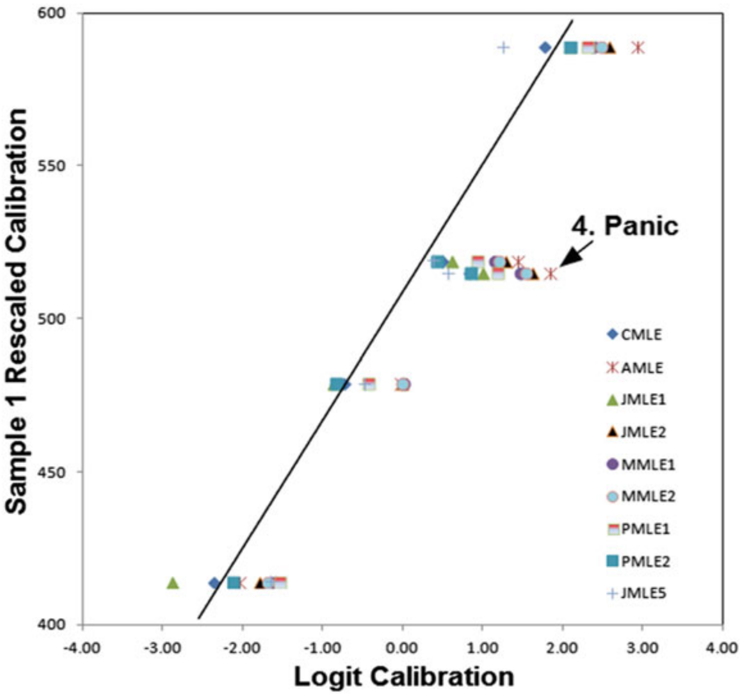


Fig. 7.4 Plot of rescaled symptom calibrations for Sample 1 against logit symptom severity calibrations for Sample 2 for shared symptoms. The trend line for CMLE is shown. Symptom 4, "Panic", is irregular and omitted from the trend line calculation

1 and Sample 2, so it is omitted from the trend line calculation. The trend line for CMLE is shown in Fig. 7.4. Its slope and intercept enable us to transform the CMLE logit calibrations into rescaled values conforming to the Sample 1 scale. The conversion equation for CMLE is: CMLE rescaled calibration = 41.4 * logit calibration + 509.1.

The conversion coefficients and the rescaled CMLE values are shown in Table 7.4. The same procedure is followed for the other sets of logit calibrations in Fig. 7.4 and are also shown in Table 7.4. In addition, the symptom score is also adapted in a similar way and its adapted values are listed in Table 7.4. An alternative rescaling procedure, outlined in Humphry and Andrich [10], is to equate the means and variances of the shared symptom calibrations. The procedure followed here was chosen because it emphasizes the importance of symptom selection, i.e., quality control, in the rescaling process.

In Table 7.4, notice that although the JMLE1 and JMLE2 rescaled values are almost identical, as are also the MMLE1 and MMLE2 results, the conversion coefficients are different. Different implementations of even the same estimation method construct different local measurement scales. The "Average of 6" column in Table 7.4 is the average of the six rescaled calibrations to its left. These differ by 2 units or less (well within their approximately 10-unit uncertainties) from the Average of 6 value despite the differences in conversion coefficients.

The values in Table 7.4 are plotted in Fig. 7.5, but with the Average of 6 calibration replacing its 6 components which would otherwise overlay on the plot. On the y-axis, the "Average of 6" value in Table 7.4 has substituted for the Sample 1 calibration for symptoms in Sample 2 not included in Sample 1.

In Fig. 7.5, some patterns emerge. The adapted Score, symbol "*", coincides with the Average of 6, symbol "⍰", at the bottom of the plot, but diverges at the top of the plot (ringed). This reflects the departure of the score-to-measure ogive from a straight line. On the plot, calibrations, with an infinite range, are linear, so scores, with a bounded range, must be curvilinear. Reassuringly, JMLE5 diverges noticeably from the Average of 6 for only two symptoms (symbol "●" ringed). Dichotomizing the 5-category rating scale has not conspicuously changed the meaning of the calibrations.

One estimation method, PMLE, is slightly out of step with the other estimation methods, and its two implementations, symbols "■" and "▲", somewhat disagree. The greatest divergence is at the top of the plot (ringed) where PMLE2 accords more closely with the Adapted Score than with the other Rasch calibrations. However, PMLE1 also outlies conspicuously near the bottom of the plot (ringed). An explanation could be that for most Rasch estimation methods, a symptom's total score is the sufficient statistic for its symptom severity calibration. For PMLE, however, the sufficient statistics for the symptoms are the sums of counts of instances of clients who score in one category on one symptom and in another category on another symptom. A consequence is that, even for complete data where every client responds to every symptom, individual responses are used in the PMLE estimation process with different frequencies. This can bias the symptom estimates in an uneven way, meaning that PMLE is the least dependable of the methods discussed here.

Table 7.4 Comparison of rescaled symptom severity calibrations for Sample 2

		Rescaled calibration										
No.	Adapted score‡	CMLE	AMLE	JMLE1	JMLE2	MMLE1	MMLE2	Avg of 6	PMLE1	PMLE2	JMLE5	Sample 1 calibration
3	409	411	410	410	410	411	411	411	419	418	410	414
4	548	543	544	544	543	543	543	543	538	539	540	515*
5	481	478	480	480	480	480	480	479	(467)	471	480	479
10	536	529	530	530	530	529	529	530	527	522	530	519
11	471	469	471	471	471	471	471	471†	(460)	467	(454)	
12	443	443	446	446	446	447	447	446†	(427)	(436)	443	
13	(590)	616	614	614	614	618	617	616†	612	(595)	(625)	
14	575	582	581	581	581	582	582	581	587	590	581	589
	−2.49	41.4	34.4	34.4	39.1	43.1	41.1		43.6	40.7	58.7	= Multiplier
	605.3	509.1	480.6	509.5	480.5	479.6	479.70		486.50	504.7	507.9	= Addition

Note: Conversion: Rescaled calibration = Multiplier * Logit calibration + Addition
Rescaled uncertainties range from 7 to 16
‡ = Score adapted by changing the dichotomous ratings from 0, 1 to 6.05, 3.57
† = Sample 2 Average of 6 Calibrations included in the Sample 1+2 combined scale
(. . .) indicates estimates 10 or more rescaled units from the Average of 6
* = Sample 1 Calibration not used in the rescaling conversion

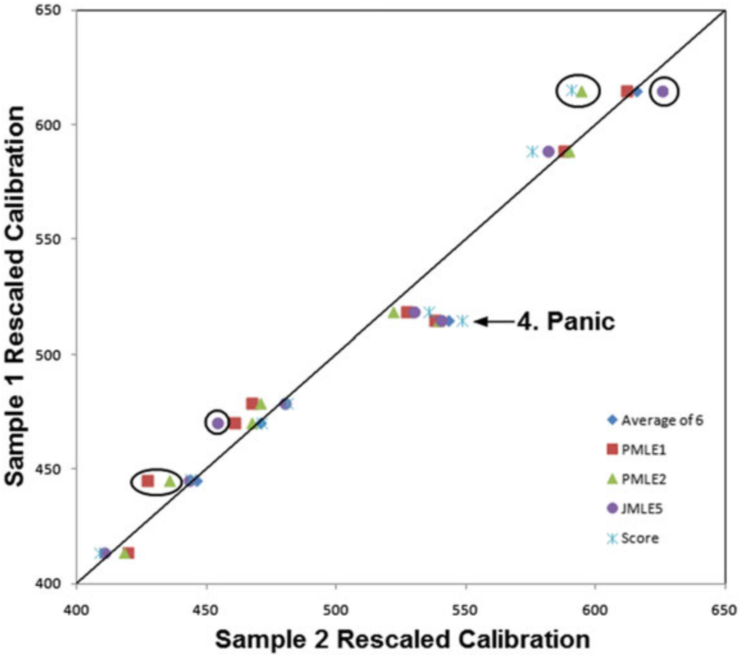


Fig. 7.5 The Sample 1 rescaled calibrations of the shared symptoms plotted against the Sample 2 rescaled calibrations. For the 3 symptoms with no Sample 1 calibration, the "Average of 6" Sample 2 calibration substitutes. The identity line is shown. Uncertainties on the x-axis are about 10 units

Choppin [6] furthermore notes that misfit in the data can also bias PMLE calibrations, reminding us of the requirement to perform quality-control of the data, such as will be done with the mean-square statistics in Table 7.6. Other estimation methods and implementations are effectively equally dependable after rescaling, so convenience can guide the choice of method and implementation in practical situations.

There is a proviso. Some software implementations offer more than one option for the calculation of client intensity measures. The standard option, though not always available, is the maximum likelihood estimate. Another option is Warm's mean likelihood estimate [41]. Further options include Bayesian adjustments [3, 39]. These need to be evaluated in a manner similar to Tables 7.3 and 7.4 to discover which client measures are generalizable with the implementation's symptom conversion coefficients. This may also alter the choice of software implementation.

7.2.4 Equating Sample 3 of Clients

For Samples 1 and 2, the data are complete. We may even think that adapted raw scores are a good enough basis for a quasi-linear measurement system. However, raw scores require locally complete data. In contrast to this inefficient one-size-fits-all approach, tailoring the phobia questionnaire to the client's situation may enable a shorter, but almost equally informative, questionnaire to be administered [4, 18, 26, 29].

We mimic this situation with Sample 3, a separate sample of 100 clients. For these clients we have a pool of 14 symptoms, indicated in Table 7.1 by "R", "B", "M", "T". Two questions, labeled "R", are administered to all 100 clients. These are routing questions for a Flexilevel test [24]. If the client responds 1 or 2 on the 5-category rating scale to both of these symptoms, then subset B of the symptom items is administered. If the client responds 3, 4, or 5 to both symptoms, then subset T of symptoms is administered. Otherwise subset M is administered.

Subsets B, M and T are each 4 symptoms, so each client responds to 6 symptoms, selected according to the client's phobia intensity. When this selection is applied to the 100 clients, 43 are administered subset B of symptoms, 31 have subset M, and 26 subset T. These data are analyzed by JMLE (using "Winsteps"). For comparison, the complete data for all 14 of these symptoms for the same 100 clients are also analyzed with JMLE (also using "Winsteps").

Table 7.5 shows the results. Its rows are sorted by the Flexilevel Logit Calibration. Scores are disordered, and so are no longer a possible surrogate for client measures because missing data compromises comparability. The symptom calibrations in Table 7.5 are plotted in Fig. 7.6. Outlying symptoms 14 and 15 (ringed) are omitted from the conversion of the logits to the rescaled calibrations. The rescaled calibrations are plotted in Fig. 7.7. The symptom calibrations for the complete 14-symptom data approximate the Flexilevel symptom calibrations reasonably and more closely than the Sample 1+2 calibrations, as should be expected given that the Flexilevel and 14-symptom calibrations are derived from the same sample's data.

The purpose behind the Flexilevel approach is to reduce the response burden on the clients without unduly reducing the measurement effectiveness of the instrument. Having established that the adaptively reduced data set defines the same phobia scale as the complete data set, we can ask whether the client measures are statistically identical across the two instrument administration methods.

Figure 7.8 plots the rescaled Flexilevel client intensity measures against their 14-symptom intensity measures. The reduction from 14 symptoms to 6 symptoms for each client shows a loss of some precision, but the trend is very strong, with most pairs of measures falling near the identity line.

Figure 7.9 shows the effectiveness of the Flexilevel routing, with 82% of clients routed correctly. The left vertical arrows at about 430 mark the top of the low range of scale values targeted for the B (both initial items rated low) condition. The right vertical arrows at about 520 mark the top of the middle range of scale values targeted for the M condition (mixed low and high ratings on the first two items). The T condition (both initial items rated high) extends from 520 to the top of the scale.

Table 7.5 Rasch severities of phobia symptoms for Sample 3

	Flexilevel symptoms (Adaptive simulation)				14 Symptoms (Complete Data)					
Symptom number	Score	Count	Logit calibration‡	Rescaled calibration	Score	Count	Logit calibration	Rescaled calibration	Sample 1+2 calibration	One word
2	116	43	−2.66	395	345	100	−1.75	397	406	Disgust
3	104	43	−2.24	414	327	100	−1.48	414	414	Uneasy
11	78	43	−1.21	460	277	100	−0.77	461	471	Skin
5	91	31	−0.77	479	256	100	−0.47	481	479	Anxious
15	68	43	−0.7	482	229	100	−0.07	507	554 *	Chills
1	254	100	−0.54	489	254	100	−0.44	483	466	Freaked
8	76	31	−0.07	510 †	220	100	0.07	516		Crazy
7	93	26	0.11	518 †	229	100	−0.07	507		Nervous
6	69	31	0.26	525 †	222	100	0.04	514		Nauseous
4	207	100	0.43	532	207	100	0.29	531	515	Panic
17	53	31	1.17	556	180	100	0.77	562	562	Shiver
13	63	26	1.82	594	153	100	1.37	601	616	Crying
9	57	26	2.17	609 †	165	100	1.08	582		Urge
14	56	26	2.24	613	151	100	1.43	605	514 *	Vomit

Flexilevel conversion: Rescaled Calibration = 44.3 * Logit + 513.3. 14-Symptom conversion: Rescaled Calibration = 65.6 * Logit + 511.5
Rescaled uncertainties for Flexilevel are 6 to 12 units, for 14-Symptoms are 8 to 11 units. * = Sample 1+2 Calibrations not used in the rescaling conversion. † = Sample 3 Calibrations included in the combined scale. ‡ = this Table is sorted by Flexilevel Logit Calibration

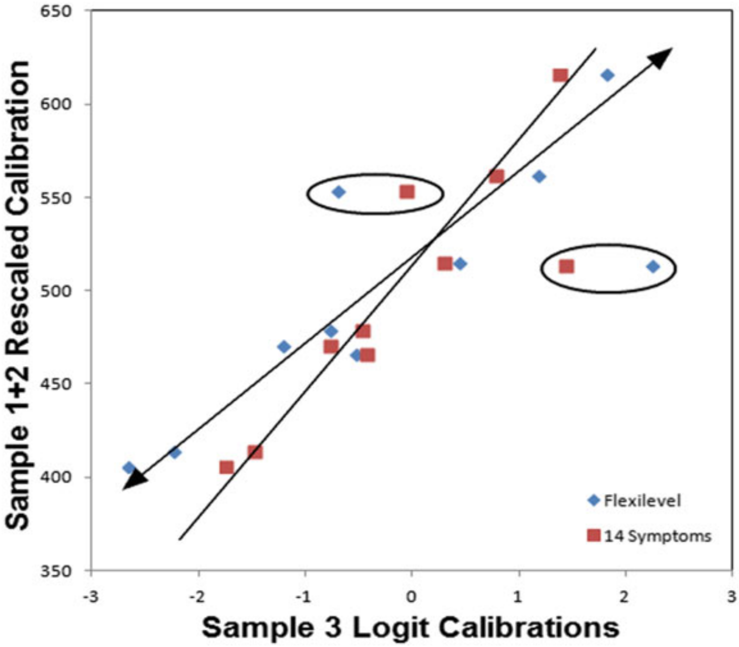


Fig. 7.6 Plot of rescaled symptom calibrations for Sample 1+2 against Sample 3 Flexilevel and 14-symptom logit severity calibrations for Sample 3. Trend lines for Flexilevel (arrowed) and 14-Symptoms are shown. Ringed estimates for symptoms 14 and 15 are far from the trend lines and omitted for their calculations. Uncertainty on the y-axis is about 5 units and on the x-axis is about .2 logits

Only 18% of the clients were routed to a less targeted set of symptoms, none by more than one level. That is, five clients in the B range have measures over 430, and five at the M level have measures lower than 430. Four more at the M level have measures higher than 520, above the intended range, and four at the T level have measures lower than 520, below the expected upper range. The Flexilevel process has accomplished its objective. Clients have not been burdened with the need to respond to questions irrelevant to their conditions, and this has been done without compromising the comparability of the resulting measurements.

7.2.5 Virtual Equating

In Fig. 7.4 the rescaling conversion was performed mathematically. We can also do it visually, a process called "Virtual Equating," which is an application of substantive variable mapping [35, 36] and construct mapping [42] methods introduced by Wright and Stone [44] and Wright and Masters [43]. This approach may be attractive to those who think visually, or who desire to work with closer involvement in the qualitative features of the measured construct.

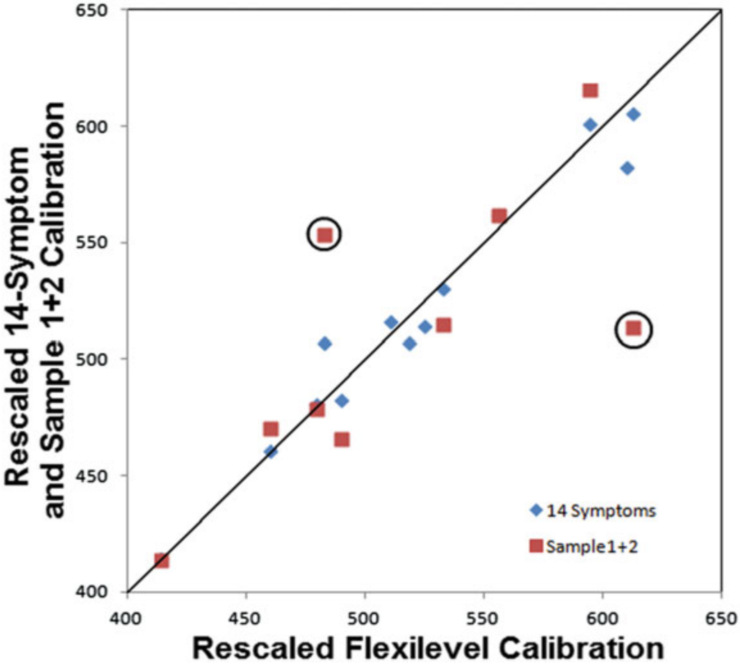


Fig. 7.7 Plot of rescaled symptom calibrations for Sample 3. The rescaled calibrations for all 14 symptoms and for Sample 1+2 are plotted against the rescaled Flexilevel symptom calibrations. The ringed points are for symptoms 14 and 15. The diagonal line is the identity line. Uncertainty on both axes is about 5 units

Figure 7.10 shows the 10 symptoms from Sample 1 at their rescaled values. Next to them are the 8 symptoms from Sample 2 positioned on their own logit phobia scale from the JMLE1 analysis. Five of the symptoms are the same, so I have used these to align the Sample 1 and Sample 2 phobia scales by eye. Keep in mind that the positions of these symptoms on the scale are estimated from the client responses and are not in any way a consequence of analyst inputs or researcher influence. As discussed above, merely numeric differences in the estimates obtained from separate samples may conceal the presence of statistically identical patterns of structural invariance. The boxes around four of the pairs of item names indicate the symptoms used for the virtual equating (Vomit, Itchiness, Anxious, Uneasy). The ringed symptom (Panic) is not used because it is too different (for unexplained reasons that may have to do with differences between the two samples of clients). This symptom is idiosyncratic in some way, so it is dropped from the virtual equating.

The end result of laying out the symptoms on the scales calibrated by the separate samples in Fig. 7.10 is confirmation of the conversion in Table 7.5. In Fig. 7.10, symptoms for Sample 2 that were not included in the Sample 1 scale have been aligned with the Sample 1 symptoms (arrowed). The combined phobia scale for Samples 1+2 is shown in the last column of Fig. 7.10.

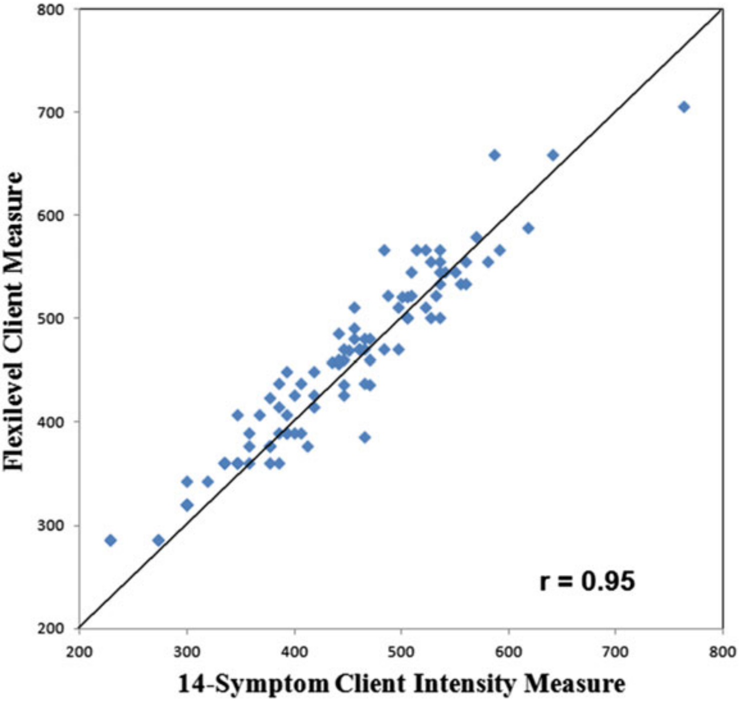


Fig. 7.8 Scatter plot of 100 client calibrations for Sample 3. The scatter reflects the loss of precision when reducing from 14 symptoms to 6 symptoms for measuring each client. Uncertainty on the x-axis is 6 to 12 units and on the y-axis is 8 to 11 units

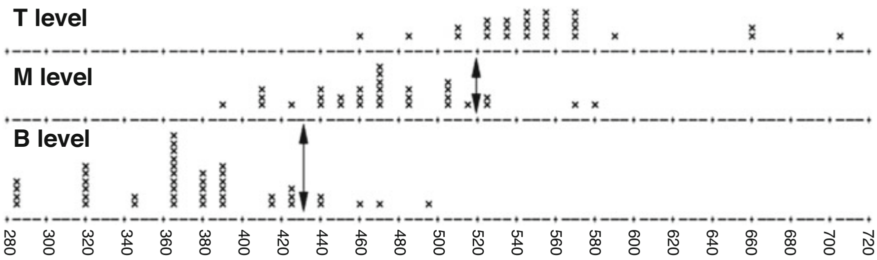


Fig. 7.9 Plot of Flexilevel client calibrations for Sample 3 showing the impact of the two routing symptoms

7.2.6 Confirming the Phobia Scale

The 17 rescaled symptom estimates have now been placed onto a combined scale locating together in a shared frame of reference the ten symptoms from Table 7.2, the three symptoms from Table 7.4, and four symptoms from Table 7.5. Let's call this combined scale the "Benchmark Phobia Scale". It comprises three subsets of symptoms which were analyzed with different estimation methods, implemented with different software, and which used different scoring mechanisms.

The intention behind the construction of the Benchmark Phobia Scale is that the scale surmounts the specific symptoms chosen, the specific estimation method or software used, and the specific format of the responses to the questions. This intention does nothing but follow through to its logical and practical consequences the decision to employ an identified measurement model [34] requiring the persistent display of structural invariances across samples and instruments.

Let us check how well we have accomplished this. We have the datasets for Samples 1, 2, 3 containing different symptoms and samples of clients. Two datasets, Samples 1 and 3, have 5-category rating scales, but we have allowed each dataset to define the measurement characteristics of its own rating scale. Sample 2 has a 2-category, dichotomous, rating scale. We will now analyze all three datasets together using JMLE implemented in "Facets" [19]. The combined dataset is thus 17 symptoms and 200 + 100 + 100 = 400 clients. "Facets" allows the same symptom to be analyzed with different rating scales. In this analysis, the Rasch Grouped Rating Scale Model (GRSM, [25]) is used. Samples 1 and 3 are modeled to share the same 5-category rating scale. Sample 2 is modeled to use its own dichotomous rating scale.

Table 7.6 lists the Benchmark Phobia Scale calibrations, shown as the "Bench Scale" column, obtained by analysis of Samples 1, 2 and 3 separately and extracted from Tables 7.2, 7.4 and 7.5 in which all the calibrations had been rescaled to accord with Sample 1. The "Combined Samples" column contains the calibrations from the "Facets" analysis of all three Samples combined, rescaled using the method of Fig. 7.4. The 17 symptoms and three samples were extracted from the complete original Trypophobia dataset. This original dataset of 17 symptoms and 582 clients is analyzed with JMLE using "Winsteps". Its calibrations, rescaled in the same way, are shown in the "Original Dataset" column. These are plotted in Fig. 7.11.

The very high correlations shown in Table 7.6 and reasonable dispersions around the identity line in Fig. 7.11 (relative to their uncertainties of roughly 10 units) are reassuring. They indicate that a usefully definitive Benchmark Scale has already been achieved. The Original Dataset contains 8976 ratings for non-extreme clients. In the three Samples, there are 3400 ratings, i.e., 38% of the Original Dataset. Despite the deliberate manipulation and abbreviation of the original data--manipulations undertaken with the goal of demonstrating how the calibrated construct might persistently display its characteristic identity across variations in samples, instruments, and estimation--the original measurement system has been preserved in such way that it can be generalized beyond these datasets.

In Table 7.6, the mean-square fit statistics for the Combined Samples analysis are all acceptable. Each mean-square is its chi-squared statistic divided by its degrees of freedom, so that its expectation is 1.0. Mean-squares in the range 0.5 – 1.5 are usually acceptable [22]. In the Original Dataset analysis, only one mean-square is alarming, 2.32 for Symptom 9, "Urge". This indicates there is a major source of substantive inconsistency in the data for this symptom. Since the mean-square for

Table 7.6 Comparison of symptom severity calibrations for different datasets

Symptom number	Bench scale	Combined Samples			Original dataset			One word
		Logit	Rescaled	MnSq	Logit	Rescaled	MnSq	
1	466	−0.79	477	1.20	−0.50	483	1.08	Freaked
2	406	−2.27	410	0.78	−1.75	408	0.88	Disgust
3	414	−2.12	417	0.83	−1.68	412	0.80	Uneasy
4	515	0.31	526	0.95	0.33	532	0.83	Panic
5	479	−0.69	481	0.87	−0.52	481	0.84	Anxious
6	525	0.20	522	1.08	0.02	514	0.74	Nauseous
7	518	−0.07	509	1.19	0.05	516	0.87	Nervous
8	510	−0.16	505	0.73	0.00	513	0.79	Crazy
9	609	2.08	606	1.20	0.97	571*	2.32*	Urge
10	519	0.29	526	1.28	0.24	527	1.40	Itchiness
11	471	−1.07	464	0.67	−0.90	459	0.82	Skin
12	446	−1.81	431	0.75	−0.83	463	0.93	Anxious
13	616	1.99	602	0.87	1.96	630	0.89	Itchiness
14	589	1.85	596	0.90	1.40	597	0.77	Vomit
15	514	0.06	515	1.19	−0.26	497	1.01	Chills
16	554	1.02	558	0.70	0.63	550	0.86	Breath
17	562	1.18	566	0.79	0.83	562	0.72	Shiver
Conversion Multiplier =			45.0			60.0		
Addition =			512.5			512.6		
Correlation with Benchmark Scale			r = 0.99			r = 0.97	* = unexpected value	
Mean Uncertainty =			7.1			3.9		

this symptom in the Combined Samples analysis is reasonable, this statistic must be associated with responses in a subset of client ratings that was not part of the three samples.

Further, this symptom is the only conspicuous outlier in Fig. 7.11. In Table 7.6 and Fig. 7.12, the Benchmark severity of Symptom 9, Urge, is 609, but the Original severity is 571, 38 units, (almost four uncertainty ranges) less. In the anomalous subset of ratings, there is a greater propensity to rate "Urge" highly than in our data subsamples. In fact, investigation of the analytic details reveals that the 17 of the 50 most unexpected ratings in the Original analysis are for the symptom "Urge". In every case, "Urge" has been rated much higher than expected. Though professionally informed opinions on spe-cific client cases may lead to insights as to why these inconsistent observations occurred, these results suggest that "Urge" is not a stable symptom of phobia, and should be removed from the Benchmark Scale.

In summary, there are already indications that improvements can be made to the choice of symptoms and to the severity values on the benchmark scale. Quality control, such as this, is an essential and continuing feature of any measurement system.

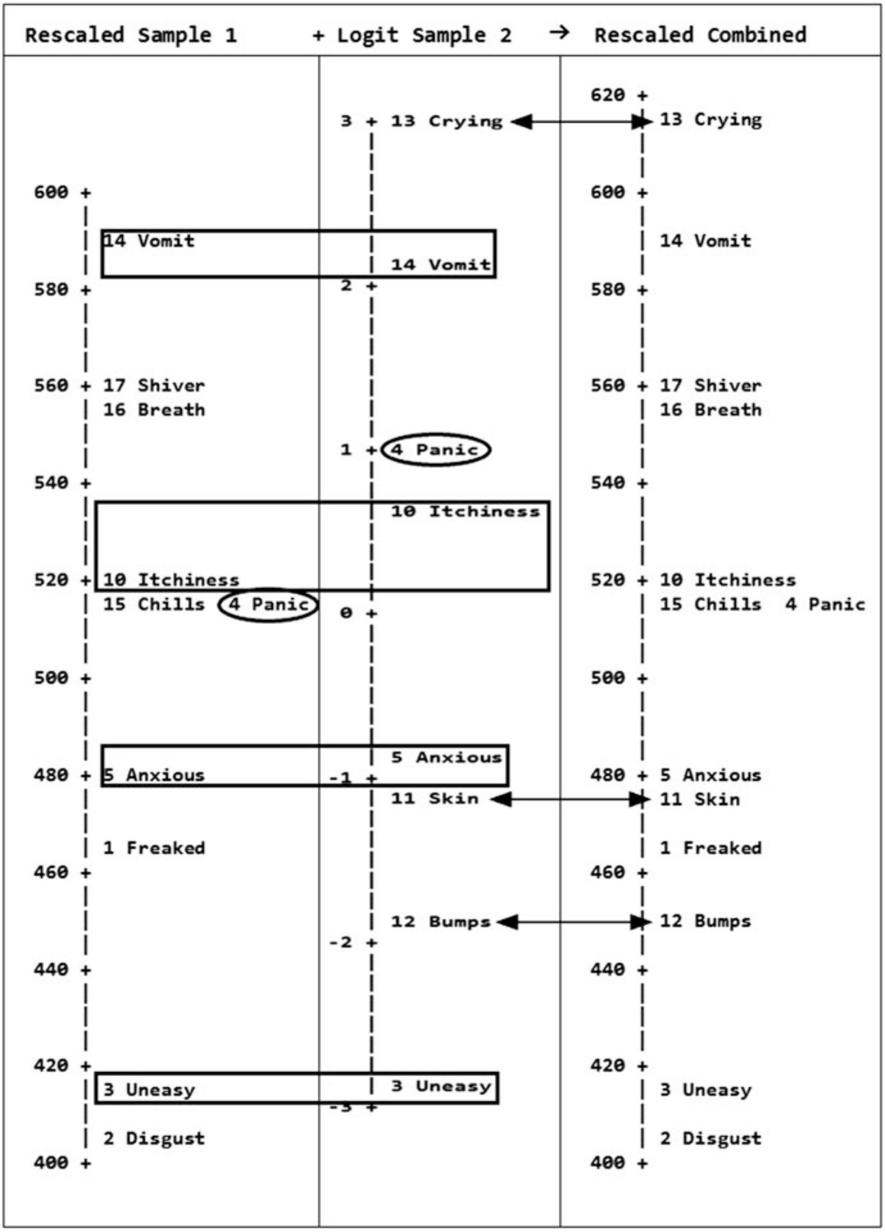


Fig. 7.10 The 10 symptoms for Sample 1 with at their rescaled calibrations and the 8 symptoms for Sample 2 with their logit calibrations equated onto one scale

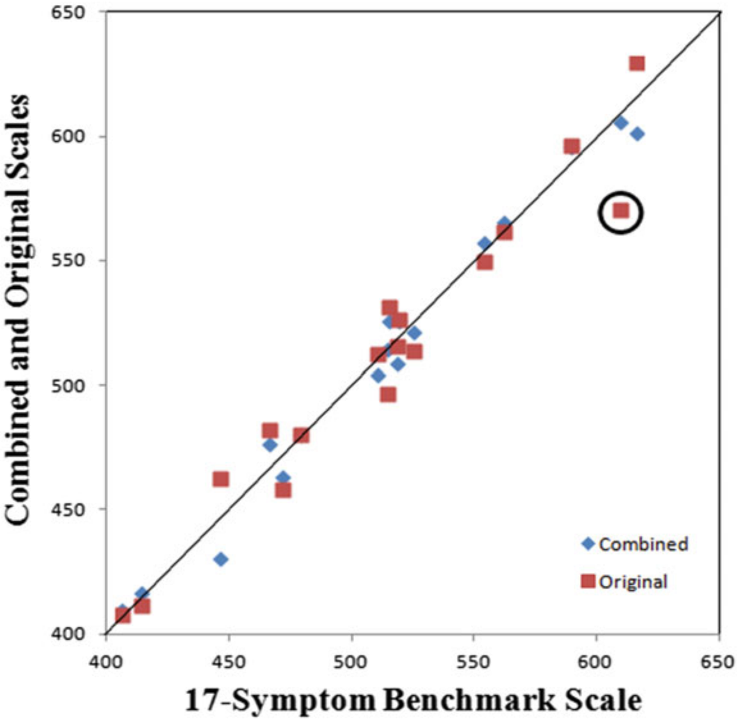


Fig. 7.11 The Combined and Original calibrations plotted against the Benchmark Scale calibrations. The identity line is shown. The ringed outlier is Symptom 9, "Urge". Uncertainty on the x-axis is 4 to 16 units and on the y-axis is 4 to 7 units

7.3 Measuring with the Benchmark Scale

Suppose that we have a new instrument containing 5 of the symptoms and a newly defined rating scale. As has been demonstrated, calibrations can be obtained from any subset of the 17 symptoms of the Benchmark Scale and further symptoms can be added, temporarily or permanently. We want to put the new 5-symptom instrument on the Benchmark scale. Here are some procedures:

1. Administer the new instrument to 20 or more clients, preferably 50 [15], then follow the procedure depicted in Tables 7.3 and 7.4. Analyze these data with your chosen software. Output logits and obtain the conversion coefficients to convert local logits onto the Benchmark scale.

If a client responds to all 5 questions, then ratings can be summed. Use the logit values for the symptoms and the Andrich thresholds of the rating scale. Then apply AMLE to each possible raw score. If the chosen software does not support this directly, then AMLE can be implemented in Excel [17]. Finally apply the conversion

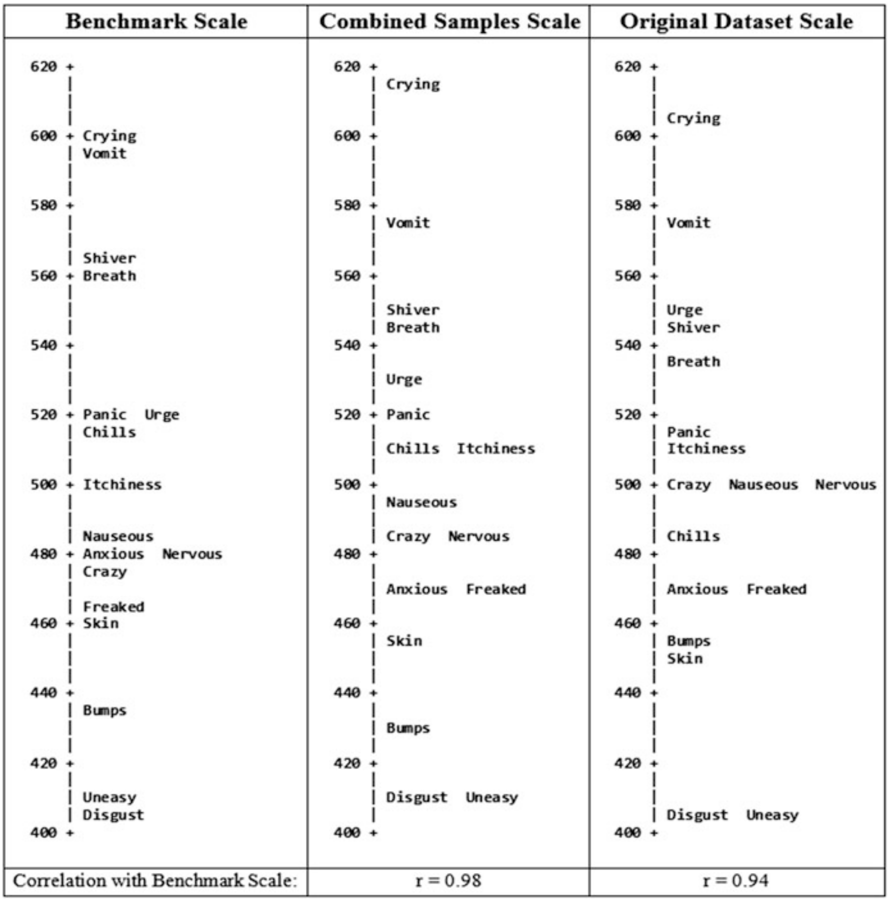


Fig. 7.12 The symptom severities from Table 7.6 shown graphically

coefficients to obtain a score-to-measure table similar to Table 7.7. If a client responds to some or all of the 5 symptom questions, the client's measure can also be computed immediately by applying AMLE only to those symptoms and the raw score. This can be implemented in a smartphone app or such like.

2. If a score-to-measure table similar to Table 7.7 is needed immediately, before any clients have responded, then it can be constructed by choosing reasonable conversion coefficients from Table 7.4 or similar. Convert the benchmark scale values for the chosen symptoms and rating from, say, Table 7.4 and Fig. 7.2, back to logits, then apply the AMLE procedure in 1 above.

Table 7.7 Client scores and Benchmark scale intensity calibrations for 5 phobia symptoms

Score	Calibration	Uncertainty	Score	Calibration	Uncertainty
5	222	112	16	536	32
6	300	64	17	552	32
7	350	48	18	570	33
8	383	41	19	588	34
9	409	38	20	608	35
10	431	35	21	630	38
11	450	34	22	656	41
12	469	33	23	688	48
13	486	32	24	739	64
14	503	32	25	817	112
15	519	31			

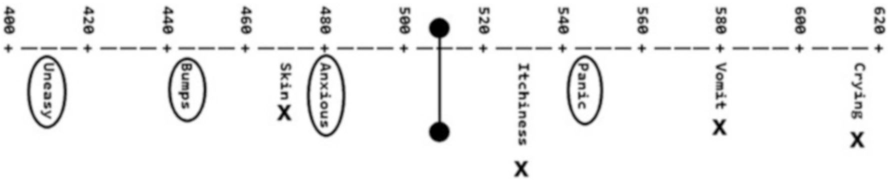
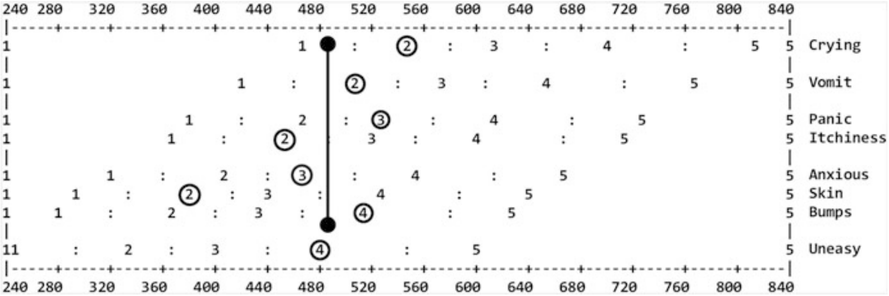


Fig. 7.13 Measurement "Keyform" for phobia based on 8 dichotomized symptoms. The vertical line shows the measurement for a client who affirmed the ringed ratings. Uncertainty in the client measures is about 35 units

7.3.1 A Phobia Intensity "Ruler"

Table 7.4 and Fig. 7.2 further enable us to construct a measurement ruler, known as a "Keyform" [14, 16], for phobia intensity. A keyform for the 8 symptoms scaled in Sample 2 is shown in Fig. 7.13. This corresponds to the dichotomized rating scale as analyzed by JMLE1 in Table 7.4. The symptoms are positioned at their severities along the Benchmark phobia scale. "O" indicates the client affirmed this symptom. "X" indicates the client did not affirm this symptom. The approximate client intensity measure is shown by the vertical line.

Figure 7.14 shows the same data with the 5-category rating scale as analyzed by JMLE5 in Table 7.4. The x-axis is the Phobia Intensity Scale. The y-axis lists the symptoms. Along each row, the rating-scale categories for each symptom are positioned at their calibrations on the Phobia Intensity Scale. Categories 2, 3, 4 for each symptom are positioned at their locations of maximum probability as shown in Fig. 7.2. The maximum probabilities for categories 1, "Not at all", and 5, "Extreme", are at opposite infinities. This is impractical, so the numerals categories 1 and 5 are positioned where the expected scores on each symptom are 1.25 and 4.75, respectively, on the 1–5 rating scale. These values are chosen such that their locations on



```
240  280    320    360    400    440    480    520    560    600    640    680    720    760    800    840
|-----+------+------+------+------+------+------+------+------+------+------+------+------+------+-------|
1                                       1      :      2     :    3      :        4         :        5    5 Crying
|                                                                                                        |
1                              1      :        2    :     3     :       4          :        5            5 Vomit
|                                                                                                        |
1                       1      :        2     :   3      :       4          :        5                   5 Panic
1                     1      :       2           3     :       4           :       5                     5 Itchiness
|                                                                                                        |
1             1      :       2     :    3      :       4         :         5                             5 Anxious
1        1      :       2     :    3      :       4          :        5                                  5 Skin
1      1      :       2     :     3     :       4           :       5                                    5 Bumps
|                                                                                                        |
11       :      2     :     3      :      4           :        5                                         5 Uneasy
|-----+------+------+------+------+------+------+------+------+------+------+------+------+------+-------|
240  280    320    360    400    440    480    520    560    600    640    680    720    760    800    840
```

Fig. 7.14 Measurement "Keyform" for phobia based on 8 5-category symptoms. The x-axis shows the client phobia intensity calibrations. For each symptom, the expected intensity calibration for each category is shown by the category number. The vertical line shows the measurement for a client with the ringed ratings. Uncertainty in the client measures is about 25 units

the Phobia Scale are distant from the intensity measures for categories 2 and 4, but not unreasonably far away. The numerals 1 and 5 at the extremes of each row remind us that those ratings continue out to infinity.

Figure 7.14 can be used immediately as a measuring device. In this example, the rings around the client's ratings allows the therapist to determine by eye a visual sense of a client's measure. A vertical line drawn through the middle of the ratings approximates a measure of the intensity of the client's phobia. Uncertainty values or a 95% confidence interval define the range outside of which unexpected responses are outliers, horizontally, such as "2" for "Skin". These exceptions may be useful for further diagnosis of the client's condition.

Notice that the precision of the estimated measure for this client on the 5-category rating scale is higher than for the dichotomized rating scale in Fig. 7.13, so that smaller changes in phobia intensity can be tracked on the 5-category scale than on the dichotomous scale. The diagnostic power (quality control) of the 5-category scale is also higher.

7.4 Discussion

In this chapter we have seen how social science variables can be constructed, blended and extended, and then refined into a useable measurement system. Using different segments of the Trypophobia dataset, we have demonstrated how the findings from different samples of clients and different subsets of symptoms can be combined to construct a comprehensive measurement scale. In the course of this, we discovered that the measurement scale can surmount the specific response structure employed to communicate with the clients. We also showed that the measurement scale accommodates different estimation methods and different software implementations of those estimation methods without distorting the meaning of the unit quantity.

There is a paradoxical finding in Table 7.4. Two estimation methods, CMLE and PMLE, are considered to be of higher quality according to statistical theory [2, 40] and in practical application [23]. They accordingly might be expected to agree with each other, and to disagree with all the rest of the estimation methods. However, after rescaling, CMLE agrees with all the other estimation methods; only PMLE does not agree, and it does not seem to provide superior estimates.

Following the procedures described here, another Phobia Questionnaire with its dataset of responses can be aligned with the Benchmark Scale, perhaps augmenting the Benchmark Scale. For a brand-new questionnaire, the graphical depiction of the Benchmark Scale in Fig. 7.14 can be used to obtain linear client measures immediately. Then, after 20 or more relevant clients have responded, a more precise alignment between the new questionnaire and the Benchmark scale can be made.

The linear Phobia measures are ideal for statistical analysis and for communication between therapists, researchers, and clients. Investigations into the causes of phobias and their treatment becomes independent of the specifics of the phobias and the devices used to collect data about them. One more unit on the Benchmark phobia scale means the same amount of change, regardless of the intensity of phobia experienced. If results of different studies using different instruments measuring the same thing were aligned and expressed in a common language, social science might at last be positioned to facilitate the kind of rapid advances the physical sciences have demonstrated are possible when we are able to learn quickly and effectively from each other's efforts.

References

1. D. Andrich, A rating formulation for ordered response categories. Psychometrika **43**, 561–573 (1978)
2. D. Andrich, Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. Psychometrika **75**, 292–308 (2010)
3. N.H. Azizan, Z.B. Mahmud, R. Adzhar, Rasch measurement model: a review of Bayesian estimation for estimating the person and item parameters. J. Phys. Confer. Ser. **1366**, 012105 (2019). https://doi.org/10.1088/1742-6596/1366/1/012105
4. M. Barney, W.P. Fisher Jr., Adaptive measurement and assessment. Ann. Rev. Organ. Psychol. Organ. Behav. **3**, 469–490 (2016)
5. T.-W. Chien, J.M. Linacre, W.-C. Wang, Examining student ability using KIDMAP fit statistics of Rasch analysis in Excel, in *Communications in Computer and Information Science: Vol. 201. Advances in Information Technology and Education, CSE 2011 Qingdao, China Proceedings, Part I*, ed. by H. Tan, M. Zhou, (Springer, Berlin, 2011), pp. 578–585
6. B. Choppin, A fully conditional estimation procedure for Rasch model parameters. Eval. Educ. **9**(1), 29–42 (1985)
7. O.D. Duncan, M. Stenbeck, Panels and cohorts: Design and model in the study of voting turnout, in *Sociological Methodology 1988*, ed. by C. C. Clogg, (American Sociological Association, 1988), pp. 1–35
8. W.P. Fisher Jr., Measurements toward a future SI: On the longstanding existence of metrology-ready precision quantities in psychology and the social sciences, in *SMSI 2020 Proceedings*, ed. by G. Gerlach, K.-D. Sommer, (AMA Service GmbH, Wunstorf, 2020), pp. 38–39. https://www.smsi-conference.com/assets/Uploads/e-Booklet-SMSI-2020-Proceedings.pdf

9. J. Heine. *Pairwise: Rasch Model Parameters by Pairwise Algorithm*. R package version 0.4.4-7 (2020). https://CRAN.R-project.org/package=pairwise
10. S.M. Humphry, D. Andrich, Understanding the unit in the Rasch model. J. Appl. Measurement **9**(3), 249–264 (2008)
11. S. Imaizumi, Y. Tanno, Rasch analysis of the Trypophobia Questionnaire. BMC Res. Notes **11**(1), 128 (2018). https://doi.org/10.1186/s13104-018-3245-5
12. E. Kennedy, *A Cultural History of the French Revolution* (Yale University Press, 1989), pp. 77–78
13. R.C. Kessler, W.T. Chiu, O. Demler, E.E. Walters, Prevalence, severity, and comorbidity of twelve-month DSM-IV disorders in the national comorbidity survey replication (NCS-R). Arch. Gen. Psychiatr. **62**(6), 617–627 (2005). https://doi.org/10.1001/archpsyc.62.6.617
14. G. Kielhofner, L. Dobria, K. Forsyth, S. Basu, The construction of keyforms for obtaining instantaneous measures from the occupational performance history interview ratings scales. OTJR **25**(1), 23–32 (2005)
15. J.M. Linacre, Sample size and item calibration stability. Rasch Measurement Trans. **7**(4), 328 (1994). https://www.rasch.org/rmt/rmt74m.htm
16. J.M. Linacre, Instantaneous measurement and diagnosis. Phys. Med. Rehabil. State Art Rev. **11**(2), 315–324 (1997). https://www.rasch.org/memo60.htm
17. J.M. Linacre, Estimating Rasch measures with known polytomous (or rating scale) item difficulties: Anchored Maximum Likelihood Estimation (AMLE). Rasch Measurement Trans. **12**(2), 638 (1998). https://www.rasch.org/rmt/rmt122q.htm
18. J.M. Linacre, Computer-adaptive testing: A methodology whose time has come, in *Development of Computerized Middle School Achievement Tests [in Korean]*, MESA Research Memorandum No. 69, ed. by S. Chae, U. Kang, E. Jeon, J. M. Linacre, (Komesa Press, Seoul, 2000). In English: http://www.rasch.org/memo69.pdf
19. J.M. Linacre, *Facets Rasch Measurement Computer Program* (Winsteps.com, Beaverton, 2020a)
20. J.M. Linacre, *Winsteps Rasch Measurement Computer Program* (Winsteps.com, Beaverton, 2020b)
21. J.M. Linacre, B.D. Wright, The length of a logit. Rasch Measurement Trans. **3**(2), 54–55 (1989). https://www.rasch.org/rmt/rmt32b.htm
22. J.M. Linacre, B.D. Wright, Reasonable mean-square fit values. Rasch Measurement Trans. **8**(3), 370 (1994). https://www.rasch.org/rmt/rmt83b.htm
23. C.-W. Liu, W.-C. Wang, Parameter estimation in Rasch models for examinee-selected items. J. Educ. Measurement **54**(4), 518–549 (2017). https://doi.org/10.1111/jedm.12159
24. F.M. Lord, The self-scoring flexilevel test. J. Educ. Measurement **8**(3), 147–151 (1971). https://doi.org/10.1111/j.1745-3984.1971.tb00918.x
25. A. Lundgren-Nilsson, A. Tennant, Past and present issues in Rasch analysis: The functional independence measure (FimTM) revisited. J. Rehabil. Med. **43**(10), 884–891 (2011)
26. M.E. Lunz, B.A. Bergstrom, R.C. Gershon, Computer adaptive testing. Int. J. Educ. Res. **21**(6), 623–634 (1994)
27. P. Mair, R. Hatzinger, M.J. Maier. *eRm: Extended Rasch Modeling*. 1.0-1 (2020). http://erm.r-forge.r-project.org/
28. L. Pendrill, *Quality Assured Measurement: Unification Across Social and Physical Sciences* (Springer, 2019)
29. K. Pesudovs, Item banking: A generational change in patient-reported outcome measurement. Optom. Vis. Sci. **87**(4), 285–293 (2010)
30. G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests* (Danish Institute for Educational Research, Copenhagen, 1960/1980) (expanded edition (1980) with foreword and afterword by B.D. Wright. The University of Chicago Press)
31. D. Rizopoulos, ltm: An R package for latent variable modelling and item response theory analyses. J. Stat. Softw. **17**(5), 1–25 (2006). http://www.jstatsoft.org/v17/i05/

32. A. Robitzsch. *Sirt: Supplementary Item Response Theory Models*. R package version 3.9-4 (2020). https://CRAN.R-project.org/package=sirt
33. A. Robitzsch, T. Kiefer, M. Wu. *TAM: Test Analysis Modules*. R package version 3.3-10 (2019). https://CRAN.R-project.org/package=TAM
34. E. San Martin, J. Gonzalez, F. Tuerlinckx, Identified parameters, parameters of interest, and their relationships. Measurement Interdiscip. Res. Perspect. **7**(2), 97–105 (2009)
35. M.H. Stone, Substantive scale construction. J. Appl. Measurement **4**, 282–297 (2003)
36. M.H. Stone, B. Wright, A.J. Stenner, Mapping variables. J. Outcome Measurement **3**(4), 308-322. [http://jampress.org/JOM_V3N4.pdf] (1999)
37. E.L. Thorndike, *The Thorndike Scale for Handwriting of Children* (Bureau of Publications – Teachers College, Columbia University, New York, NY, 1912)
38. L.L. Thurstone, Psychology as a quantitative rational science. Science **LXXXV**, 228–232 (1937). (Rpt. in L. L. Thurstone. (1959). The measurement of values (Midway Reprint Series) (pp. 3–11). University of Chicago Press)
39. J.S. Uebersax, Statistical modeling of expert ratings on medical treatment appropriateness. J. Am. Stat. Assoc. **88**, 421–427 (1993)
40. M. Von Davier, Rasch model, in *Handbook of Item Response Theory*, ed. by W. van der Linden, (Routledge, 2016)
41. T.A. Warm, Weighted likelihood estimation of ability in item response theory. Psychometrika **54**, 427–450 (1989)
42. M.R. Wilson, *Constructing Measures: An Item Response Modeling Approach* (Lawrence Erlbaum Associates, 2005)
43. B.D. Wright, G.N. Masters, *Rating Scale Analysis* (MESA Press, 1982)
44. B.D. Wright, M.H. Stone, *Best Test Design* (MESA Press, 1979)
45. B.D. Wright, R.J. Mead, L.H. Ludlow, *KIDMAP: Person-by-Item Interaction Mapping*, Tech. Rep. No. MESA Memorandum #29 (MESA Press, 1980). http://www.rasch.org/memo29.pdf

Chapter 8
Equating Measuring Instruments in the Social Sciences: Applying Measurement Principles of the Natural Sciences

David Andrich and Dragana Surla

Abstract The concept of measurement in which the magnitude of a property is quantified in a common unit relative to a specified origin is a deep abstraction. This chapter shows the application of measurement in a social science context where the motivation is transparency and equity rather than the advancement of scientific laws. However, to achieve these, the realization of measurement needs to be no less rigorous than it is in the advancement of scientific laws. Rasch measurement theory provides the basis for such rigor. The context in this chapter is competitive selection into universities in Western Australia based on a summary performance on a series of instruments which assess achievement in a range of discipline areas. Such selection tends to determine life opportunities; therefore to ensure consistency and fairness, performances on different instruments need to be transformed into measurements which are in the same, explicit unit relative to a specified origin. Because the illustrative context is complex, it is considered that the Rasch measurement theory applied in this chapter could be applied to a range of social contexts where assessments on different instruments need to be transformed to measurements in a common unit referenced to a common origin and where the focus is on making decisions at the person level.

Keywords Instrument equating · Quantification · Theory, Rasch models, Measurement

D. Andrich (✉) · D. Surla
The University of Western Australia, Perth, Australia
e-mail: david.andrich@uwa.edu.au

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3_8

8.1 Introduction

Measurement, the quantification of the magnitude of some property of an object from a specified, convenient or natural *origin* in a constant *unit* of an instrument, is a deep abstraction. For example, the elementary measurement of mass using a beam balance is both simple and sophisticated. Though the balance is a relatively simple instrument, the *conceptualization* to quantify the mass of the object in a meaningful way is a remarkable abstraction. The balance involves having a three-dimensional object, potentially of any shape, volume, color, and so on, on one side, and a set of equivalent objects on the other side that balances it, mapping the count of this set on a real number line, itself partitioned into equal contiguous distances. There is little in the natural world that even approximates a real number line, which is totally abstract, where even a drawn line to represent it is ragged when looked at through a microscope.

Despite aspects of deep abstraction and scientific implications, school-children understand the concept of measurement of mass using a beam balance, and understand more generally the idea of measurement in a constant unit relative to a specified origin. The motivation for measurement of common properties was not to advance physical laws, but the fair trade of objects with properties such as mass, length, and volume. The standardization of units for this purpose is exemplified by the development of the metric system [1]. Although standardized for purposes of fair trade, the metric system was developed by scientists of the highest order. The relationship of science to measurement from this perspective is summarized by Alder:

> We often hear that science is a revolutionary force that imposes radical new ideas in human history. But science also emerges from within human history, reshaping ordinary actions, some so habitual we hardly notice them. Measurement is one of our most ordinary actions. We speak its language whenever we exchange precise information or trade objects with exactitude. This very ubiquity, however, makes measurement invisible. To do their job, standards must operate as a set of shared assumptions, the unexamined background against which we make agreements and make distinctions. So it is not surprising that we take measurement for granted and consider it banal. Yet the use a society makes of its measures expresses its sense of fair dealing. That is why the balance scale is a widespread symbol of justice [1, p. 1].

This chapter is concerned with scientific measurement in which the prime purpose is transparency and fairness, and not the advancement of laws. Nevertheless, for this important purpose, measurement needs to be as rigorous as that needed to advance quantitative, scientific laws, and needs to be of the same kind that advances such laws. Thus, although rigorous measurement will be required for the development of quantitative laws in the social sciences, there are social contexts where transparency and fairness of decisions seem sufficient to require rigorous scientific measurement. This chapter provides such an example.

8.1.1 A Complex Social Science Context

In educational assessment of proficiency, it is common to refer to the instruments of assessment as *tests*. Because the example of this chapter is illustrative of a general approach to equating, and because of the connections made with measurement in the natural sciences, they are referred to as *instruments*. The typical study and process of equating scores from different instruments assumes that they assess the same variable [24]. The context of the example of this chapter, referred to as the *frame of reference* for reasons that emerge later, is substantially more complicated than that. The context is the selection of students into universities in Western Australia based on their assessment on *instruments* from an array of *disciplines* at the end of 12 years of schooling. There are some 40 disciplines of possible study, and students must have scores on at least four, including in the discipline of English, though many have scores on five or more disciplines. Although there are prerequisites for university entry in some fields, such as engineering, to meet other policy requirements – for example, not specializing studies too early – these are minimal. In addition, where there are prerequisites, additional electives may be chosen, and these are not the same among the candidates. Each instrument's scores range from 0 to 100 and, because a summary score of each student's profile is calculated and the students are ranked for competitive entry into universities in Western Australia, the scores of all instruments need to be equated onto a measurement scale with the same unit relative to a specified origin.

Within each discipline area, the instrument scores can be considered to reflect *causal* variables in which proficiency of the student in the discipline area governs the performance on the instrument. In principle, within each instrument, different items that assess the same proficiency are exchangeable. However, the summary score across a range of disciplines, can be considered an index variable [4, 39, 40, 43]. This variable can be considered a higher order, *thick* variable, one thicker than those from each of the disciplines, that reflects a general capacity to profit from a university education as evidenced from previous relevant study.

The frame of reference is complex for the following reasons. First, the summary index variable is even more complex than the kind illustrated by Stenner, Burdick and Stone [39]. Their example defines socioeconomic status (SES) by *education, occupational prestige, income, and neighborhood.* In defining SES in this way, none of the components are exchangeable. These same components define the variable for each person. However, in the example of university selection, they are not defined by a fixed set of disciplines, and in principle, they are exchangeable. For example, two candidates may be competing for entrance to the same university study, such as law or psychology, with only one instrument score out of four or five which is common. Second, because candidates self-select the disciplines they study, and they have scores on different combinations of instruments, no pair of instruments have entirely common candidates. Third, although the instrument scores are positively correlated, the correlations among instrument scores are not homogeneous. For example, the correlation between instruments of Mathematics and Chemistry is greater than that between either discipline and English or History. Fourth, the scores are probabilistic, not deterministic, relative to proficiency, and for their analysis a measurement theory that is inherently probabilistic is required. Finally, as indicated above, for various historical reasons, the scores have a finite range, and therefore, especially close to the higher limits of the range where competitive scores are most relevant, the relationships among the scores are not linear. How each of these complexities is accommodated in producing measurements on the same scale relevant for the purpose of university entry is the substance of this chapter. Because of the complexities of this frame of reference, it is considered that many other contexts that require equating of instruments might be accommodated by the application of the Rasch measurement theory described in this chapter.

8.1.2 Empirical Understanding and the Role of the Rasch Model for Measurement

In the natural sciences, advanced measurement is derived from the scientific, theoretical understanding of the relevant variables and their relationships [26]. Direct reading of measurements from the instrument hides the substantive theory and design that *manifests* the property measured and *controls* properties that disturb it. The beam balance exploits the effect of gravity and simply controls other factors, for example measurements in very small units of chemical properties used antique balances that were enclosed to ensure no disturbance from air movement. Appropriately anchored, and elegantly for such an elementary instrument, the balance will give the same measurement whether under the gravitational force of the Earth or the Moon and whether it is stationary or accelerating relative to either of them. From the perspective of scientific theory, the separation of the concepts of mass and gravity took the geniuses of Galileo and Newton, and at the time was controversial. The studies of mass and gravity in physics, from nuclear energy to gravity waves, continue to be advanced areas of physics. However, to ensure transparency and equity, everyday transactions also require accurate measurement of mass.

The beam balance is not only familiar, but its application is relatively tangible, thus disguising some of the sophistication in the understanding of mass and gravity. On the other hand, the now equally familiar mercury thermometer for measuring human and day temperatures, though now easy to apply, is conceptually much less tangible than the beam balance. In particular, the origin and unit are more abstract than a unit of mass, for example, a kilogram. In Celsius, at one atmosphere of pressure, 0° C is set at the freezing point of water, 100° C at the boiling point, and the unit is one 100th of this range. These end-points are chosen for convenience of relatively observable everyday phenomena. Empirical work was required to ensure that the uniform expansion of mercury in a thin tube in the range specified, also implied uniform increases of temperature.

The thermometer requires the control of the relative expansion of all of its other components. This control requires the scientific understanding of vacuums, pressure of gases, and so on. This is the reason that the construction of reliable, practical thermometers required the work of the best scientists of the nineteenth century. Indeed measurement of temperature was developing before the concept of heat as the kinetic energy of atoms was understood, meaning that the measurement of this variable and its understanding were more or less simultaneous [15]. An important inference from this example is that the reliable measurement of a variable indicates a substantial understanding of its properties, and reciprocally, that successful measurement can enhance further understanding.

In the frame of reference in the example of this chapter, the counterpart to the theoretical understandings of variables are the transparent, explicit syllabuses for each discipline for Years 11 and 12 of schooling, the teaching of these syllabuses by qualified teachers, the design of instruments to assess a range of proficiencies that reflect each syllabus, and the anonymous scoring of the performances of each candidate by two independent qualified markers. Even the studies up to Year 10, in the development of student proficiencies to enable them to study at Years 11 and 12, are relevant. Thus a great deal of professional, intellectual and empirical work goes into the production of scores on the instruments by students. The ultimate, substantive, validity of each instrument is the public outcry in the newspapers, and these days, social media, if questions stray from the syllabus.

It is stressed that such substantive, empirical work is essential and that the application of Rasch measurement theory to equate instruments and to map observed scores into measurements cannot overcome, through probabilistic modelling, any shoddy, superficial, and poorly constructed instruments that do not validly reflect candidates' relative proficiencies in the respective disciplines. The function of the application of Rasch measurement theory, where scores of all instruments are valid, is to ensure that the different instruments are transformed onto a measuring instrument with the same unit relative to the same specified origin. The reason this is necessary is that each instrument is designed to align the range of the difficulties of its items to the proficiencies of the candidates, and that relative to the higher order index variable described above, these proficiencies are not the same in each discipline. In addition, the structure, format, and scoring of different questions, which is natural to the different disciplines, is not the same across these disciplines. As a result, every instrument has its own implied relative unit, which may not even be consistent across its own continuum of proficiency. Finally, each is referenced to its own relative origin; in common parlance, instruments are more or less difficult and instrument scores are more or less skewed.

8.1.3 Descriptions of Measurement

Because of its role in science and everyday applications, it is not surprising that measurement has been studied by physical and social scientists, among many of them are Campbell [13], Duncan [16], Finkelstein [19], Fisher and Stenner [20], Luce and Tukey [27], Mari [30], Ramsay [35], and Wright [47]. Perhaps surprisingly, physical scientists can, and most do, take the concept and need for measurement for granted. On the other hand, social scientists generally do not have that luxury. This section does not provide a review of discourses on the history, structure, function and definitions of measurement; instead, to set up the distinctive definition in Rasch measurement theory, and why this definition is most germane to the example of social science measurement of this chapter, it only summarizes briefly three common definitions of measurement.

Measurement assumes that magnitudes of properties can be mapped on a real number line. This is assumed with the measurement of mass and temperature summarized above. With this assumption, three definitions are most common. First *classical*, second *representational*, and third *additive conjoint*. They all attempt to describe measurement by formalizing the properties illustrated above with the measurement of mass and temperature.

The classical definition emphasizes measurement as the ratio of the amount of the property of an object relative to an amount defined as the unit [33]. The representational definition emphasizes that the relationship among the properties is the same as the relationship among the numbers assigned to the objects [25]. For example, for certain physical relationships among masses, concatenating two objects is equivalent to having a single object whose mass is the sum of the measures of each object. The additive conjoint definition is more abstract, requiring that the rows, columns, and cells of a two-way table with real numbers, which reflect the magnitudes of properties, can be transformed monotonically to produce an additive structure [27].

Rather than considering some of its operational aspects as a basis for formalizing a definition of measurement, Rasch arrived at his requirement from his empirical studies in reading proficiency [36]. Specifically, that within a *specified frame of reference*, and with stimuli (referred to as instruments in this chapter) and individuals characterized by real numbers,

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison… Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; … [37, p. 332].

From the mathematical abstraction of his definition, and further derivations from them, Rasch observed that the properties of measurement that are characterized by classical, representational, and additive conjoint, are satisfied [36]. However, he considered that the requirement of *invariant comparisons* stood as a more fundamental basis for measurement than any *description* of measurement [38]. Andrich [6] makes the case that, not only is Rasch's definition compatible with the other three, but that it *explains* them. Moreover, the other definitions are deterministic, whereas Rasch's theory is set in both deterministic and probabilistic contexts, where the probability characterizes the uncertainty in the observation when one person encounters one instrument. It is not concerned with the distributional properties of populations of persons. This chapter applies the probabilistic formulation.

8.1.4 Measurement of Variables with No Physical Counterpart

Rasch's formulation of invariant comparisons, conveniently and elegantly, abstracts measurement further than typically considered in the natural sciences in that it makes no reference to any physical properties. Rasch was not the first to consider such an abstraction and the quest for invariance of comparisons. In abstracting measurement from his work on comparative judgements of pairs of a set of objects with respect to a physical property such as sound pitch or luminescence, Thurstone commented:

> One of the main requirements of a truly subjective metric is that it shall be entirely independent of all physical phenomena. In freeing ourselves completely from physical measurement, we are also free to experiment with aesthetic objects and with many other types of stimuli to which there does not correspond any known physical measurement [44, pp. l82–83].

Thurstone emphasized that the mapping of the magnitude of the property on a real number line involved very specific focus on a variable, and the control of all other variables. In the construction of instruments for measuring attitudes in terms of statements that reflected different degrees of the attitude through the opinion they expressed, he writes that

> The various opinions cannot be completely described merely as "more" or "less". They scatter in many dimensions, but the very idea of measurement implies a linear continuum of some sort, such a length, price, volume, weight and age. When the idea of measurement is applied to scholastic achievement, for example, it is necessary to force the qualitative variations into a scholastic linear scale of some kind [44, pp. 218–19].

On the requirement of invariance of comparisons, he articulates that

> If a scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help to construct it. This may turn out to be a severe test in practice, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale [44, p. 228].

Rasch's requirements of invariant comparisons with respect to any properties that can be characterized by real numbers, physical or not, are compatible with Thurstone's. However, the distinctive part of Rasch's formulation is that the requirements are expressed formally in mathematical terms [37, 38]. As a result, further mathematical derivations can be carried out. Among other epistemological implications of these derivations [6], it makes it possible to apply the resultant model to real data. This chapter is concerned with one such example where the measurement model applied to define a unit and origin of a standard instrument could only have been formulated through a mathematical derivation.

8.1.5 Structure of This Chapter

The rest of this chapter is structured as follows. Section 8.2 provides a summary of the polytomous Rasch model and the special case of the *Rasch distribution* which makes tangible the analogy between measurement in the natural sciences and Rasch measurement theory. The thresholds in the Rasch distribution are equidistant, and the common distance is identical in interpretation to the unit of a measuring instrument in the natural sciences. In addition, this distribution, which is the inferred distribution of replications, is a discrete analogue of the continuous Gauss distribution of uncertainty for replicated measurements. In order to satisfy the unidimensionality property of the Rasch model, and therefore optimize the relationship between the instruments in deriving the equating functions, this section also provides a rationale for editing the profiles of persons. This editing is based on obtaining relatively homogeneous profiles, that is, those for which the total score is sufficient. In a complementary fashion, the section also provides a rationale for identifying, at the person measurement stage, those profiles which can and cannot be characterized by their total score. Then, because of the frame of reference of the measurement, those profiles that cannot be summarized by their respective total scores need to be considered in terms of both a summary estimate and the properties of the profile.

Section 8.3 provides an empirical example from the frame of reference described above. Specifically, it shows the details of real data from six instruments used for university entrance examinations and the results of equating these instruments to a common instrument of the Rasch distribution, where the origin and unit are defined identically, and not merely analogously, to those of an instrument measuring physical variables and are chosen for convenience in their frame of reference. Descriptive statistics before and after equating are provided with the example, emphasizing its illustrative properties that can be transferred to other contexts. The final section is a summary.

8.2 The Rasch Model and Distribution

From three successive papers [2, 3, 37], where the theoretical characteristic of these papers is emphasized by there being no data analysis in any of them, the Rasch model for ordered categorical data can be expressed in the now familiar forms, as either a *rating* or *partial credit* parameterization [31] according to

$$\mathrm{P}\{X_{ni} = x; \beta_n, \psi_{xi}\} = [\exp(\psi_{xi} + x\beta_n)]/\gamma_{ni}, \tag{8.1}$$

where X_{ni} takes integer values $x = 0, 1, 2, \ldots, m_i$ when person n, measure β_n, is assessed with instrument i; $\psi_{xi} = -\sum_{k=0}^{x}\tau_{ki}\, k = 1, 2, \ldots .m_i$, τ_{ki} are the instrument's m_i thresholds where $\tau_{0i} \equiv 0$ is introduced for notational convenience; and γ_{ni} is a normalizing factor. In common psychometric applications, the instrument in

Eq. (8.1) is an item of an instrument or questionnaire. The relevant part of the partial credit parameterization, relative to the rating parameterization, is that in the former all the thresholds of the different instruments can have different values, whereas in the latter the means of the thresholds of the instruments can be different but the deviations of the thresholds from their mean is the same across instruments.

The response structure for any one person responding to one instrument is the same in both parameterizations. However, interpreting the thresholds as *steps,* and presenting Thurstone thresholds reconstructed from the Rasch model [31, 32], is not compatible with the Rasch model [5, 34]. Moreover, although *estimates* from data for an instrument can result in reversed threshold values relative to their natural order, it is evidence of a problem with the responses produced by the instrument. Although there are multiple reasons for this conclusion, as seen later, there are two particular reasons thresholds must be in their natural order for the purposes of this chapter. First, if they were disordered, it would not be possible to define a unit of an instrument as the common, equal distance between successive thresholds, as in a measuring instrument in the physical sciences and in this chapter. Second, the distribution of uncertainty around any measurement would be bimodal, whereas all random error distributions of uncertainty, including the Gauss distribution are not only strictly unimodal, but the transition between successive probabilities is smooth. With thresholds in their correct order, the Rasch model satisfies this criterion of smooth, strict unimodality [9].

In any analysis, the parameter β is expressed in what are commonly referred to as *logits*, though the logit is not a unit of the instrument of the kind found in the natural sciences, nor of the unit as defined in this chapter [22]. Specifically, the same logit value across separate analyses of different data sets is not expressed in the same unit in the sense of the unit used in the present chapter.

At *threshold* τ_{ki}, the probabilities of responses in its two adjacent categories are equal. With a maximum score of 100 for each instrument, as in the example of this paper, there are 100 thresholds. This is an over-parameterization of the model, and with zero frequencies in the data, especially with low scores, direct estimation fails without modifying the model in some way [28, 46]. Therefore, instead of attempting to estimate all thresholds directly, they are estimated through their first four *principal components* given by

$$\begin{aligned}\psi_{xi} = \quad & -x\delta_i + \{x(m_i - x)/2\}\Delta_i + \{x(m_i - x)(2m_i - x)\}\lambda_i \\ & +\left\{x(m_i - x)\left(5x^2 - 5xm_i + m_i^2 + 1\right)\right\}\zeta_i,\end{aligned} \tag{8.2}$$

where δ, Δ, λ, ζ characterize the location, spread, skewness and kurtosis of the thresholds of the instrument. It is stressed that these parameters characterize properties *of the thresholds*, *not* the distribution of persons. It is also emphasized that the person parameter β is a scalar, and therefore is said to be *unidimensional* [7].

It is relevant to stress the implied interpretation of the Rasch model of Eq. (8.1). Namely, that given the values of the instrument parameters, then for a given value of β, Eq. (8.1) is the inferred distribution of responses as if the same person responded to the same instrument an infinite number of times. Clearly, this is not administratively feasible, but even if it were, if the same person responded to the same instrument multiple times, there would be substantial local dependence. However, that is not the point; as part of the abstraction from the data by applying the model, the distribution is an *inferred distribution* of replicated responses. An important aspect of this distribution when applied to real data is that, by analogy to the Gauss distribution, it is a *random* distribution [9]. This implies that no unaccounted-for factors are producing systematic errors in the measurements, which is checked by both ensuring fit between the data and the model and that the thresholds estimates are in their natural order [10].

8.2.1 The Expected Value Curve and the Equating Function

The estimation of the instrument parameters is considered briefly in a later section. Here we consider the relationship between observed scores and person estimates given an instrument's values (δ_i, Δ_i, λ_i, ζ_i), which may be estimates. The estimate of $\widehat{\beta}_n$ for each person is given by a maximum likelihood estimate (MLE) individually. However, the relationship between a score x_{ni} and the estimate is analytic, and holds whether or not any person in a sample obtained such a score.

For a set of instruments for which person n has scores, $\widehat{\beta}_n$ is given by the solution to the implicit equation

$$t_n = \sum\nolimits_{i_n=1}^{I_n} x_{ni} = \sum\nolimits_{i_n=1}^{I_n} E[X_{ni}] = \sum\nolimits_{i_n=1}^{I_n} \sum\nolimits_{x=0}^{m_i} x\widehat{P}\mathrm{r}\{x_{ni}\}, \tag{8.3}$$

where i_n, I_n indicate the instruments to which person n has responded and t_n is the total score on these instruments. Thus, for the person estimates β to be on the same scale, not all persons need to respond to all instruments, a key feature of the application of the model in the frame of reference of the example.

For each instrument, Eq. (8.3) specializes to Eq. (8.4), which given the observed score x_i, gives the estimate $\widehat{\beta}_{xi}$ on instrument i. Reciprocally, given any person value β, Eq. (8.4) also gives the expected value, $E[X_i|\beta]$, on the instrument:

$$x_i = E[X_i] = \sum\nolimits_{x=0}^{m_i} x\widehat{P}\mathrm{r}\{x_i\}. \tag{8.4}$$

Equation (8.4) can be solved for each total score x_i whether or not any person obtained that score, giving a unique estimate which we denote $\widehat{\beta}_{xi}$ and drop the person subscript n. In this case, though real numbers, because $x_i \in \{0, 1, 2, \ldots, m_i\}$ are discrete, the $\widehat{\beta}_{xi}$ are discrete. On the other hand, Eq. (8.4) can be solved for $E[X_i]$ given any β and is in principle continuous. Both relationships of Eq. (8.4) are applied in this chapter. This application is introduced in Fig. 8.1.

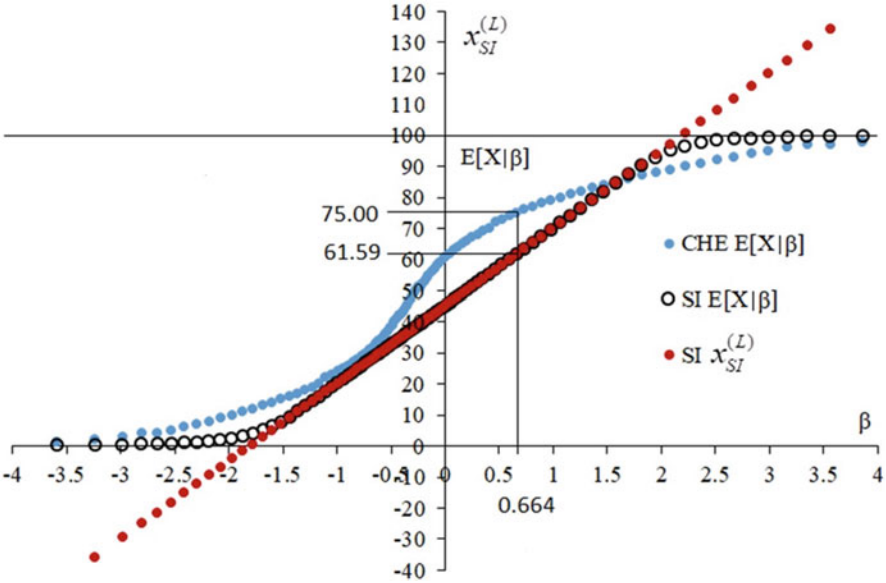


Fig. 8.1 $E[X|\beta]$, $(\delta, \Delta, \lambda, \zeta) = (-0.06325, 0.05025, 0.00024, 0.00001)$ for CHE; $E[X|\beta]$, $(\delta, \Delta) = (0.20000, 0.04000)$; and the linear measurement function $x_{SI}^{(L)}$ of the Standard Instrument (SI), $x_{SI}^{(L)} = E\left[x_{SI}^{(L)}|\beta = 0\right] + \beta_x/\Delta = 45 + 25\beta_x$

Figure 8.1 shows the relationship between the observed score x_i and person estimate $\widehat{\beta}_{xi}$, and the expected value curve $E[X_i|\beta]$ as a function of β, for one of the instruments called CHE which is analyzed in detail later in this chapter. The values of the parameters of this instrument are shown in the caption of Fig. 8.1. Because of the large value of 100 for the maximum score, the values of these parameters are small – therefore they are shown to five decimal places. Although referred to as an item characteristic curve when responses are dichotomous because it is more descriptive in its meaning, the graph of Fig. 8.1 is referred to as an *expected value curve* (EVC).

Figure 8.1 also shows a second EVC of an instrument that is linear over a substantial range of the continuum with curvature only at the extremes. This instrument has only two parameters, the first two principal components (δ_i, Δ_i) of Eq. (8.2), whose values are shown in the caption of the Figure. In anticipation of further detail below, this instrument is referred to as the Standard Instrument (SI), and is the instrument to which all instruments are equated or mapped. The SI has equidistant successive thresholds, Δ_i, the value which is identical to $\beta_x - \beta_{x-1}$ over a well-defined range. Therefore, it is analogous in interpretation to the unit of a standard measuring instrument. Moreover, again in anticipation of further elaboration, over the same range of the values, the random, uncertainty distribution is the discrete counterpart of the Gauss distribution, the distribution of random variation of replicated measurements. The SI is not simply an empirical regression equation, but an instrument whose unit and origin are both explicit and are deliberately chosen for context relevance and convenience. Accordingly, a value on the SI is referred to as a *measurement*.

Here we note that Fig. 8.1 shows the effective mapping function of the instrument CHE onto the SI. In particular, the score $x = 75$ on CHE, $\left\{\widehat{\beta}|x = 75\right\} = 0.664$, gives the measurement, $E[X_{SI}|\beta = 0.664] = 61.591$, on the SI. Finally, Fig. 8.1 shows a full linear extrapolation of the SI beyond the minimum and maximum scores of 0 and 100, which is explained later in this chapter.

8.2.2 The Rasch Distribution of Uncertainty

We now explain in detail the SI. A special case of the Rasch model of Eq. (8.2) involves just the first two principal components, (δ_i, Δ_i), giving

$$\mathrm{P}\{X_{ni} = x; \beta_n, (\delta_i, \Delta_i)\} = [-x\delta_i + \{x(m_i - x)/2\}\Delta_i + x\beta_n]/\gamma_{ni}. \tag{8.5}$$

This distribution has a distinctive role in showing measurement of the kind found in the physical sciences. Therefore, two of its characteristic features are elaborated now: first, the presence of the explicit unit Δ relative to the specified origin δ, directly equivalent to that in the physical sciences; and second, the resultant random distribution of measurement uncertainty which is the discrete analogue of the Gauss distribution of uncertainty of replicated measurements in the physical sciences.

Because of these features, Eq. (8.5) is referred to as the *Rasch distribution*, rather than simply a model.

8.2.3 The Unit in the Rasch Distribution

The EVC of the empirical instrument in Fig. 8.1 has a non-linear relationship with the continuum β. The non-linearity results from the presence of skewness and kurtosis in the thresholds. On the other hand, because its thresholds have no skewness or kurtosis, the SI is linear over a substantial range. An excerpt of the relationship between x, $E[X_{SI}|\beta]$ and β is shown in Table 8.1, where the observed scores and expected values in the range 15–86 inclusive are highlighted in bold. Table 8.1 also shows a column referred to as $x_{SI}^{(L)}$ which is defined formally below as *measurements* on the SI.

Not only is the relationship between x, $E[X_{SI}|\beta]$ and β linear in the range shown, but it makes the unit explicit. Thus the difference between two successive values β_{x+1}, β_x is not only constant and linear with the observed score x, but the difference between them is exactly the unit $\Delta = 0.04$, that is, $\beta_{x+1} - \beta_x = \Delta$; $x = 15, 16, 17,$

Table 8.1 The relationship between x, $E[X_{SI}|\beta]$, β of the Standard Instrument of Fig. 8.1

x, $E[X\|\beta]$	β	$\beta_{x+1}-\beta_x$	$x_{SI}^{(L)}$		x, $E[X\|\beta]$	β	$\beta_{x+1}-\beta_x$	$x_{SI}^{(L)}$
0	−3.114	.	−32.9		…	...	…	...
1	−2.400	0.714	−15.0		82	1.480	0.040	82.0
2	−2.051	0.349	−6.3		83	1.520	0.040	83.0
3	−1.880	0.171	−2.0		84	1.560	0.040	84.0
4	−1.768	0.112	0.8		85	1.601	0.040	85.0
5	−1.684	0.084	2.9		86	1.641	0.040	86.0
6	−1.616	0.068	4.6		87	1.682	0.041	87.1
7	−1.557	0.059	6.1		88	1.724	0.042	88.1
8	−1.504	0.053	7.4		89	1.766	0.042	89.2
9	−1.456	0.048	8.6		90	1.810	0.044	90.3
10	−1.410	0.046	9.8		91	1.856	0.046	91.4
11	−1.366	0.044	10.9		92	1.904	0.048	92.6
12	−1.324	0.042	11.9		93	1.957	0.053	93.9
13	−1.282	0.042	13.0		94	2.016	0.059	95.4
14	−1.241	0.041	14.0		95	2.084	0.068	97.1
15	−1.201	0.040	15.0		96	2.168	0.084	99.2
16	−1.160	0.040	16.0		97	2.280	0.112	102.0
17	−1.120	0.040	17.0		98	2.451	0.171	106.3
18	−1.080	0.040	18.0		99	2.800	0.349	115.0
19	−1.040	0.040	19.0		100	3.514	0.714	132.9

$SI:\delta=0.200;\ \ \Delta=0.040; x_{SI}^{(L)}=E\left[x_{SI}^{(L)}|\beta=0\right]+\beta_x/\Delta=45+25\beta_x$

…, 84, 85, 86. This relationship can be shown algebraically. Thus let the measurement on the SI be notated $x_{SI}^{(L)}$ where the superscript (L) indicates that $x_{SI}^{(L)}$ is linear throughout, and not an expected value or an observed measurement which is constrained between 0 and 100. Then relative to $E[X_{SI}|\beta=0]$,

$$\beta_x=\left\{x_{SI}^{(L)}-E[X_{SI}|\beta=0]\right\}\Delta, \tag{8.6}$$

showing that values of β_x increase by the value of the unit Δ for each integer increase in the observed score x. In addition, relative to the origin, this relationship between an observed integer count x and the value of β is directly analogous to a measurement of an object, relative to its origin, in the unit of the instrument.

Rearranging Eq. (8.6), gives the general relationship

$$x_{SI}^{(L)}=E\left[x_{SI}^{(L)}|\beta=0\right]+\beta/\Delta. \tag{8.7}$$

The last column of Table 8.1 shows values of $x_{SI}^{(L)}$. Note Eq. (8.7) is not estimated as a regression equation, but is expressed analytically as a relationship between the measurement $x_{SI}^{(L)}$ given the value of β. Table 8.1 shows that in the range in which the relationship is linear between x, $E[X_{SI}]$ and β, their values are identical to three decimal places. These values are also highlighted in bold. Outside this range, for the same value of β, $x_{SI}^{(L)}$ is different from x, $E[X_{SI}]$. Table 8.1 also shows that only at the extremes of the range of the instrument, 0, 1 and 99, 100, the values of $x_{SI}^{(L)}$ show very large extrapolations. This in part is a result of the choice of the unit and origin. How this range is determined is described in the next sub-section. However, it is because of the properties of the distribution, in particular that of Eqs. (8.6) and (8.7), the values of $x_{SI}^{(L)}$ have been, and continue to be, referred to as *measurements*.

8.2.4 The Rasch Distribution of Measurement Uncertainty

As indicated above, the Rasch model distribution of Eq. (8.1) is the *inferred distribution* of replications of responses of the same person to the same instrument. This is simply a property of the probabilistic model. However, this inference holds for data only if the responses also fit the model. If they *do* fit the model, and the thresholds are ordered, then this distribution is of random uncertainty, with no evidence that any unaccounted-for factors are disturbing the responses [10]. This same inference holds for the Rasch distribution of Eq. (8.5) which has only the two parameters, the origin and the unit of the instrument specified. Of course, in this distribution of the SI, the thresholds are defined to be ordered. Because this distribution is directly analogous to the Gauss distribution of measurement uncertainty, now taken for granted in the natural sciences, we briefly review the motivation and role of the Gauss distribution in measurement.

Although the Gauss distribution is so mainstreamed that it is referred to generally as the *normal distribution*, the motivation and lengthy evolution of this distribution is generally not presented in textbooks. Besides Gauss, the derivation of the distribution exercised the best mathematicians in the late 18th and early 19th centuries, including De Moivre, Lagrange, La Place, and others [17, 41]. It was derived to account for the consistent evidence that *…repeated measurement of a fixed quantity by the same procedure under constant conditions…* did not give the same values but a distribution of values. The derivations culminated … *in the quadratic exponential law of Gauss* [17, p. 1]. This distribution satisfied the requirement that it characterized variation that was random, it having been realized that rather than propagating errors, random variation cancelled them. Thus the distribution is a theoretical distribution of random variation of replicated measurements, not a distribution derived to describe any particular data set. However, to the degree that any data set does conform to the Gauss distribution, to that degree it provides evidence that variation is no more than random, and therefore that relevant inferences can be drawn from the data, for example, that the mean is an ideal characterization of the object of measurement. Measurement of uncertainty continues to be a concern of natural scientists [23].

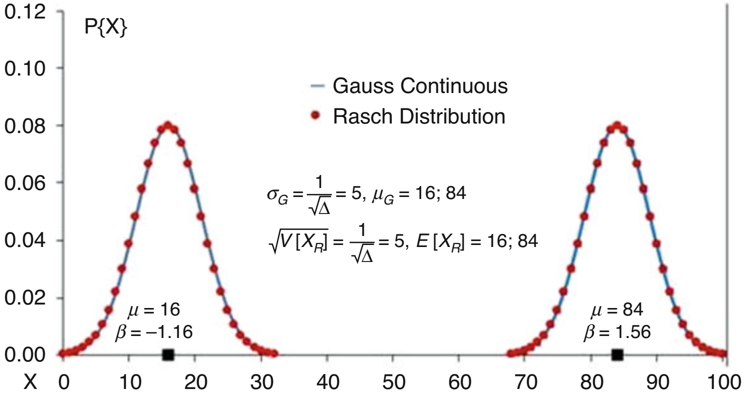


Fig. 8.2 Two Rasch distributions of the Standard Instrument where the probabilities of extreme measurements vanish, interpolated by the Gauss distribution in which $E[X_R] = \mu_G$, $V[X_R] = \sigma_G^2 = 1/\Delta$

Figure 8.2 shows the Rasch distribution for the SI of Fig. 8.1 in which $(\delta, \Delta) = (0.200, 0.040)$. The figure also shows the mean, $E[X|\beta]$ for the two measurements $(\beta_l, \beta_u) = (-1.160, 1.560)$. The Rasch distribution is clearly discrete with respect to the possible integer measurements. The discrete probabilities in Fig. 8.2 are interpolated with a continuous distribution. Perhaps unexpectedly, this is the continuous Gauss distribution. The possibility of this interpolation is no coincidence. For completeness, Eq. (8.8) shows the now common form of the Gauss distribution,

$$P\{X = x|\mu, \sigma^2\} = \left[\exp - (x - \mu)^2/2\sigma^2\right]/\sqrt{2\pi}\sigma, \tag{8.8}$$

where $(\mu,\ \sigma^2)$ are the mean and variance of the distribution.

In elaborating this relationship, we first note an observation made by Gauss regarding the limits of the applicability of the distribution of Eq. (8.8):

> Gauss commented that (1) (*the distribution*) cannot represent a law of error in full rigor because it assigns probabilities greater than zero to errors outside the range of possible errors, which in practice always has finite limits; that such a feature is unavoidable because one can never assign limits of error with absolute rigor; but this shortcoming is of no importance in the case of (1), because it "decreases so rapidly, when $[(x - \mu)^2/2\sigma^2]$ has acquired a considerable magnitude, that it can safely be considered as vanishing." [18, p. 2].

The values $(\beta_l, \beta_u) = (-1.160, 1.560)$ were chosen because they are just inside the limits of the range for the SI where the probabilities of the extreme measurements, 0 and 100, *vanish*. Thus they are the limits within which the Gauss distribution would be applicable. Four aspects of the relationship between the Rasch and Gauss distributions are relevant to note here. First, as evident from the first term in Eq. (8.2), just as is the Gauss distribution, the Rasch distribution is a *quadratic exponential*. Second, within the range in which the Gauss distribution is applicable,

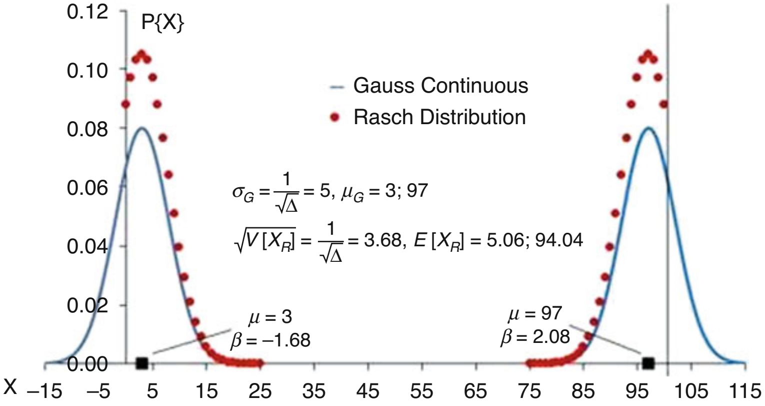


Fig. 8.3 Two Rasch distributions of the Standard Instrument where the probabilities of extreme measurements do not vanish and in which $E[X_R] \neq \mu_G$, $V[X_R] \neq \sigma_G^2 \neq 1/\Delta$

the mean of the Rasch distribution is identical to the mean of the Gauss distribution, $E[X_R] = \mu_G$, and in particular, the distribution is *symmetrical* about the mean. Third, and perhaps most surprisingly, the variance of the Gauss distribution is not only the variance of the Rasch distribution, $V[X_R] = \sigma_G^2$, but it is also the *inverse* of the unit in the Rasch distribution, $V[X_R] = \sigma_G^2 = 1/\Delta$. Finally, and importantly, it is within the range in which the Gauss distribution holds, that the relationship between the measurements β, the observed scores x, and expected values, are linear as shown in Fig. 8.1, Table 8.1 and Eq. (8.6).

For completeness, and because it is relevant to account for the finite range of the instrument, Fig. 8.3 shows the Rasch and Gauss distributions of two locations, $(\beta_l, \beta_u) = (-1.680, 2.080)$, where the limits of the instrument play a role. The figure shows the Gauss distributions as if the range of the instrument did not play a role, and the Rasch distributions which take account of the range. It is evident that the Gauss distributions are outside the range of the instrument, a feature which concerned Gauss as indicated in the above quote. The Rasch distributions of course are within the limits of the range. However, reflecting the impact of the limited range of the instrument, the mean and variance ($E[X_R]$, $V[X_R]$) of the Rasch distributions are regressed relative to their values if the range were not constrained. The property of the estimates $\widehat{\beta}_x$ is that they *undo* this regression to a large degree.

Given the properties described above, this chapter presents a justification for transforming the estimates of proficiencies β from each instrument to a measurement on the SI of the form of the Rasch distribution with the origin and unit chosen for convenience in the frame of reference. Specifically, from the estimates of the person locations β_x for each score x of an instrument, measurements in the chosen unit of the SI can be obtained from the linear extrapolation of Eq. (8.7). This linear extrapolation is shown in Fig. 8.1 and Table 8.1. It is evident that from this extrapolation, the measurements close to the limits of the range of the instrument are outside this range. However, these measurements are linear extrapolations that indicate the values that would have been obtained had the limits of the instrument not regressed the observed scores. It is these linearized measurements to which the observed scores on all instruments are mapped in the example shown in the next section.

It would be more desirable if assessments were such that few if any person locations were affected by the limits of the range of the scores (0 and 100) with these instruments, and it is generally achieved in the frame of reference of the example described next. However, where they do occur, especially at the higher limit where competition is most relevant, then the limits of the range for scores close to the extreme need to be taken into account. It is an example where the Rasch distribution, which is discrete and is a function of both the unit and maximum score on each instrument, can be applied to advantage.

8.2.5 Maximum Likelihood Estimates of the Instrument Parameters in the Rasch Model

Estimates of instrument parameters ψ_{xi} of Eq. (8.2) can be derived from Eq. (8.9) below, where $(x_{ni}, x_{nj}) \mid r_n$ are the responses of person n to two instruments (i, j), and where $r_n = x_{ni} + x_{nj}$ is the sufficient statistic for β_n:

$$\Pr\left\{ \left(x_{ni}, x_{nj}\right) \middle| r_n; \beta_n, \psi_{xi}, \psi_{xj} \right\} = \left[\exp\left(\psi_{xi}, \psi_{xj}\right) \right] / \gamma_{ni}. \tag{8.9}$$

Because of sufficiency, Eq. (8.9) is *independent* of β. To estimate parameters $(\delta_i, \Delta_i, \lambda_i, \zeta_i)$ of multiple instruments, Eq. (8.9) is generalized over multiple instruments I. The algorithm, described in detail in Andrich and Luo [7], is implemented in the software RUMM2030Plus [12] and is used in the analysis of the example of this chapter. The only constraint required in the estimation is that $\sum_{i=1}^{I} \widehat{\delta}_i = 0$. Then δ_i is the *relative* origin of each instrument in an analysis. This origin, as shown with the SI above and illustrated with the example below, can be defined independently.

Because the coefficient of each principal component is a function of frequencies in all categories, rather than of each category, the estimation is not impeded by the presence of zero frequencies, except in very extreme cases, nor is it impeded by the structurally missing data.

8.2.6 Profile Analysis and Editing of Profiles

One of the complexities of the frame of reference listed above is that the latent correlation among the instruments is not 1, which implies lack of unidimensionality. The effect on the parameter estimates of the instruments when analyzed with the Rasch model is that they are all regressed to their mean [29, 42].

To account for this feature of the data from the perspective of the frame of reference, a complementary focus to unidimensionality, one which focuses on the profiles of persons, is taken [42, 45]. Unidimensional instruments, those with a latent correlation of 1, have two implications for profiles with respect to the Rasch model: first, that each person's profile is relatively homogeneous; second, that the total score is the sufficient statistic for the person parameter, with no further information in the profile. For example, a relatively high score on one instrument implies a relative high score on other instruments, where the variation among the scores of the profile on the instruments is no more than random. In contrast, and at the other extreme, if the latent correlation were 0, there is no such relationship.

Therefore, not only from the perspective of the application of the Rasch model, but also from the requirements of the frame of reference, the profiles that need to be used to obtain the equating functions between instruments are those that are relatively homogeneous. This ensures that the properties of the instruments as reflected by their parameters, for example their relative difficulties are a property of the instruments and not of the persons who happen to have scores on the instruments. In more general terms, the profiles that need to be used are from those persons who are relatively equally proficient on all instruments.

A comparison and contrast might be made with equipercentile scaling which can be applied when instruments do not have a latent correlation of 1, but are administered to the same sample of persons [24]. Here the assumption is that, although the individual persons are not expected to have homogeneous profiles, the latent distributions of the proficiencies of the *sample as a whole* are the same for the different instruments, and any differences in scores is a property of the instruments. Therefore, with this assumption, scores with the same cumulative percentage on each instrument are deemed equivalent. This method has its own problems for equating, including with zero frequencies and extreme scores, and in addition, impeding its application in the example of this chapter, the students with scores on different pairs of instruments are not common. However, when applicable, the assumption is the equivalence of the distributions on the different instruments. It is relevant to compare the assumptions made between the equipercentile and Rasch model applications to equating. In applying the former, the assumption is that the sample has the same proficiency distribution on the different instruments; in applying the latter, it is ensured that the persons whose profiles are used have equivalent proficiencies on the different instruments.

The method of obtaining the subset of profiles whose scores are homogeneous requires two successive analyses of the data. The first is simply the standard analysis of all data. Then, given the estimates of the instruments' parameters, and each person's estimate of β, the expected value, $E[X_{ni}]$, is calculated for each instrument using Eq. (8.4). A comparison is then made between the observed score x_{ni} and $E[X_{ni}]$ for each person n on each instrument i for which they have a score. This comparison is made in terms of the standardized residual,

$$z_{ni} = (x_{ni} - E[X_{ni}])/\sqrt{V[X_{ni}]}. \tag{8.10}$$

Then if the absolute value of the residual is greater than some chosen magnitude, that score in the profile is deleted, creating further missing responses. Because there are already structurally missing responses, further missing responses are no impediment to the estimation. However, it is necessary to choose the magnitude of the residual criterion judiciously. If it is too large, there will be a substantial number of heterogeneous profiles in the analysis; and if it is too small, then relative to the variation of the model, there will be insufficient variation creating a form of local dependence. This method of editing profiles is analogous to that used by Andrich, Marais and Humphry [11] and Andrich and Marais [8] in editing responses to control the bias on item parameter estimates from guessing for multiple choice items. Specifically, given each person's and each item's parameter estimates, if there is a greater probability than random that the person guessed or at least partially guessed a response on that item, whether the response is correct or not, the response is converted to missing data. This editing of responses removes the bias in the item parameter estimates due to guessing in the data.

The criterion for deleting a response chosen in the analysis of the example in this chapter is $|z_{ni}\ | \ > 0.85$. Evidence which shows this choice is reasonable is that the latent correlation between each pair of instruments, when corrected for error, is of the order of 1. This implies that, for the purpose of obtaining equating functions, the complementary properties of unidimensionality and sufficiency of the total score of the Rasch model are satisfied.

The classical definition of reliability, when applied to the Rasch model estimates, is given by

$$r_{\beta i} = \left(V\left[\widehat{\beta}_i\right] - [V[\widehat{\varepsilon}_i])/V\left[\widehat{\beta}_i\right], \tag{8.11}$$

where $V[\beta_i]$ is the estimate of the variance of the persons on instrument i, and $V[\varepsilon_i]$ is the estimate of its error variance [21]. The latent correlation between two instruments, corrected for attenuation because of error, is given by

$$\rho_{ij} = r_{ij}/\sqrt{r_{\beta i} r_{\beta j}}. \tag{8.12}$$

In the application of Eqs. (8.11) and (8.12) in the Rasch model, $\widehat{\beta}_{xi}$ is the estimate of proficiency of the person given the score x and $V[\widehat{\varepsilon}_i]$ is the mean error variance of the estimates $\widehat{\beta}_{xi}$ from the persons who have scores on both instruments. These estimates are elaborated next.

8.2.7 Person Estimates

Given estimates $(\widehat{\psi}_{xi})$ of instrument *i*, the maximum likelihood estimate $\widehat{\beta}_n$ for each person *n* from all instruments the person has responded to, is given *individually* by the solution to implicit Eq. (8.3). However, in application to the frame of reference of this chapter, it is necessary to have an estimate, $\widehat{\beta}_{ni}$, of each person on each instrument. This estimate is given by the solution to Eq. (8.4). For this estimate of each person's proficiency on each instrument, the parameters $(\widehat{\psi}_{xi})$, following the editing of the profiles in the terms described above, are used. These are the estimates based on homogeneous profiles which, because of sufficiency of the total score for these profiles, are independent of the actual distribution of the person parameters.

However, for the person estimates based on these instrument parameters, and for evidence of sufficient proficiency for selection into university studies, every score of each person, that is the full profile before editing, must be used. Finally, each estimate of each person from each instrument is transformed to a measurement on the SI in the form of the Rasch distribution described above.

Because the original profiles are used for the final estimates, the latent correlations between instruments will not be in the range 0.90–1.00. That means that there will be profiles which, when transformed to the SI, will not be homogeneous and the sum of the estimates, or their mean, will not characterize the profile fully. How the distinction between those profiles that are homogeneous, and those that are not, is dealt with in its frame of reference is described in the context of the example. The next section provides the results of analysis of data from the example.

8.3 An Illustrative Example

The data for university selection from 2018 were provided by the School Curriculum and Standards Authority of Western Australia. As indicated above, the example comes from a series of instruments used to assess the proficiency of students in a range of disciplines for university selection in Western Australia. The total number of disciplines is as large as 40. For the purpose of illustration in the example of this chapter, scores from examinations of the following six disciplines were analyzed: English (ENG), English Literature (LIT), Mathematics 1 (MA1), Mathematics 2 (MA2), Modern History (HIM), and Chemistry (CHE) which was introduced in Fig. 8.1. These disciplines were chosen because relative properties of these instruments can be anticipated, and they illustrate some complexities that are overcome.

First, one of the disciplines of English must be taken to be eligible for university entry, and either ENG and LIT are acceptable. Therefore, very few students have scores in both disciplines. On the one hand, because it is a specialized unit, students studying LIT may be expected to have greater proficiency in English, and therefore a higher mean proficiency on the SI than those studying ENG.

Second, MA1 and MA2 have a partly different relationship from that between ENG and LIT. MA2 has more challenging material than MA1. However, to study MA2 it is necessary to either study MA1 simultaneously or otherwise know its content. Therefore, it is expected that MA2 will be shown to be more difficult than MA1, and that the mean proficiency of students studying MA2 (and MA1) will be greater on the SI than that of those studying only MA1. Finally, the disciplines HIM and CHE are chosen because one is a humanities and the other a science discipline, and they are expected to show properties that are commensurate with ENG and MA1 respectively.

For purposes of efficiency of exposition, the results are not presented in the order in which they were obtained. Following the summary of the raw data, the results of the estimates of the proficiencies and their relationships among the disciplines are shown first, followed by the estimates of the instruments' parameters, the equating functions, and finally graphical presentations of the equating functions and distributions.

8.3.1 The Raw Scores on the Instruments

An excerpt of the data file used for analysis is presented in Table 8.2, illustrating the structurally missing responses in the data file and integer scores recorded for the six disciplines.

Table 8.3 shows descriptive data in the form of the number of students, the means, standard deviations and the skewness of the distribution, and the observed pairwise frequencies and correlations. First, it is evident that ENG has the greatest number of students, which is expected because an English discipline assessment is a

Table 8.2 Excerpt of the first 15 cases from the data file for analysis

ID	CHE	ENG	HIM	LIT	MA1	MA2
S00001		45				
S00002	56	56			64	
S00003	39				50	
S00004		10				
S00005	39				50	
S00006	56	56			64	
S00007	59	61	67		43	
S00008	52				79	72
S00009			46			
S00010	58	50			35	
S00011					64	57
S00012					64	
S00013	59	61	67		43	
S00014		62			40	
S00015		80				

Table 8.3 Pairwise frequencies (F), observed correlations r_{ij} between instruments, and descriptive statistics of observed scores for each instrument from the total sample of 13617

F/r_{ij}	ENG	LIT	MA1	MA2	HIM	CHE
ENG		*	2800	937	1589	3480
LIT	*		661	267	381	698
MA1	0.289	0.247		1530	237	3255
MA2	0.409	0.348	0.867		39	1221
HIM	0.581	0.635	0.434	0.566		266
CHE	0.417	0.433	0.764	0.788	0.578	
N	10,974	1463	4426	1594	2014	4973
Mean	57.96	70.89	65.14	61.24	60.06	58.50
St dev	11.38	9.96	18.65	19.43	13.18	17.31
Skew	−0.37	−0.89	−0.65	−0.50	−0.89	−0.44

Note. *Frequency of less than 20 not shown. Number of common persons above the diagonal; observed correlations below the diagonal. Correlation between ENG and LIT not shown because of the small number of common students

requirement for university entry. LIT and MA2 have the least number, reflecting their specialist status. The table shows no common students between ENG and LIT. There were seven common students but the reason for students taking both disciplines is idiosyncratic to different circumstances and the correlation between the two disciplines was −0.188. Therefore, this frequency and the correlation are not shown in Table 8.3.

Second, all but one mean is in the range between 55 and 65, showing that the difficulties of the instruments were well aligned to the expected proficiencies of the students, except for LIT which has a mean above 70. Proficient students choose LIT, but this mean seems relatively high. There is a greater relative range in the standard deviations, where the two mathematics and the science instruments show greater standard deviations than the humanities instruments and all distributions are skewed negatively. The table also shows that the number of students assessed by each instrument varies, as does the number who are assessed by any pair of instruments. Because of the different samples, the properties of the distributions such as their means cannot be compared directly. Important from the perspective of measurement on a single variable, is that the observed initial correlations among the instruments has a large variation, ranging from 0.247 to 0.867.

The data in Table 8.3 are shown graphically in Fig. 8.4. The differences in the distributions, including their negative skewness is clear. Some of the low scores might be taken as resulting from not taking the examination seriously, but these scores are included in the analysis for completeness and illustration. Not only does the distribution of LIT show a high mean, it is also very narrow. It is also clear that the distributions of CHE, MA1 and MA2 are somewhat similar as are those of ENG and HIM. The former three instruments assess mathematics and science disciplines, the latter two assess humanities. Although the distributions cannot be compared directly, it can be inferred how well the instruments align themselves to the proficiencies of the samples. This alignment is important in distinguishing validly

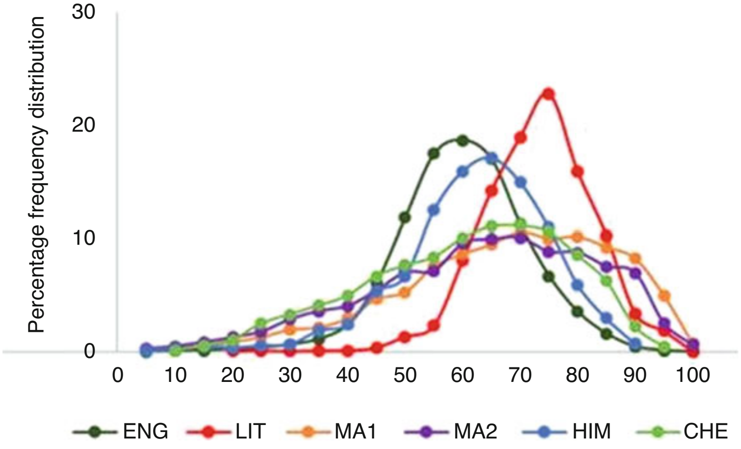


Fig. 8.4 Percentage frequency distributions of observed scores in class intervals of 5 score points

between candidates. In addition to LIT, and even though they spread the students well, the three science instruments were relatively lenient, while the two humanities instruments were better aligned. ENG, which assesses an effectively compulsory discipline in which the sample is less self-selected than the other disciplines, has noticeably the smallest value for its mode.

8.3.2 The Equating Functions

The section begins with a graphical presentation of the equating functions and the principal component estimates from which the equating functions are derived. The equating functions are obtained from an analysis in which the profiles were edited as described above. The section finishes with the latent correlations between instruments from these equating functions and the equated scores of all instruments on the SI.

Figure 8.5 shows the equating functions based on EVCs of the kind shown for CHE in Fig. 8.1. In addition, each curve has its observed means in 10 class intervals shown. It is evident that the means are very much on the curves, indicating fit to the model for the data from these instruments. An approximate Chi Statistic value which compares these observed means and their expected values across all instruments is 32.237 on 54 degrees of freedom, confirming excellent statistical fit. It is clear from Fig. 8.5 that the curves are non-linear and intersect, and require equating before comparisons can be made. This evidence that the edited data fit the model is considered sufficient for this example.

Table 8.4 shows the principal component estimates and their standard errors. Because, with a maximum score as large as 100, the values of the parameters are small in magnitude, they are shown to five decimal places. As expected, the relative

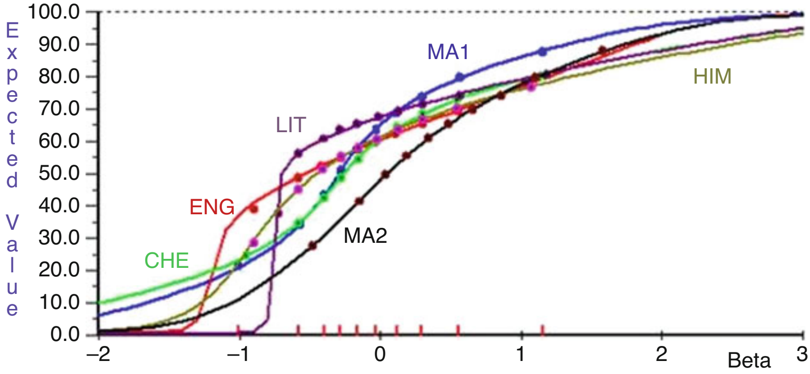


Fig. 8.5 EVCs from edited profiles with observed means in 10 class intervals

Table 8.4 Estimates of the four principal components $(\delta, \Delta, \lambda, \zeta)$ of the thresholds from a profile analysis with residuals $|z_{ni}| > 0.85$ removed

	δ(origin)	$SE(\delta)$	Δ(unit)	$SE(\Delta)$	λ(skew)	$SE(\lambda)$	ζ(kurt)	$SE(\zeta)$
ENG	−0.08586	0.00220	0.03747	0.00020	0.00055	0.00000	0.00000	0.00000
LIT	0.08213	0.00630	0.03257	0.00040	0.00110	0.00000	0.00000	0.00000
MA1	−0.23313	0.00290	0.03530	0.00020	0.00008	0.00000	0.00000	0.00000
MA2	0.19814	0.00510	0.03385	0.00020	0.00019	0.00000	0.00000	0.00000
HIM	0.10197	0.00510	0.04553	0.00040	0.00069	0.00000	0.00000	0.00000
CHE	−0.06325	0.00280	0.05025	0.00020	0.00025	0.00000	0.00000	0.00000
Mean	0.00000	0.00407	0.03916	0.00027	0.00048	0.00000	0.00000	0.00000
St Dev	0.14267	0.00150	0.00649	0.00009	0.00035	0.00000	0.00000	0.00000

difficulty of MA2, the specialist mathematics discipline, is more difficult ($\delta = 0.19814$) than MA1, the one that has the content as its prerequisite ($\delta = -0.23313$). The difficulty of LIT, the specialist English discipline, ($\delta = 0.08213$) is likewise more difficult than ENG, which is studied by most students as a required discipline, ($\delta = -0.08586$).

8.3.3 The Equated Scores to Measurements on a Standard Instrument

For each observed score of each person on each instrument, the proficiency estimate $\widehat{\beta}_{ni}$ is obtained from Eq. (8.4) and CHE is illustrated in Fig. 8.1. Then each estimate $\widehat{\beta}_{ni}$ is transformed further to a measurement on the SI according to Eq. (8.7) in which $E[X_{SI}|\beta_{SI} = 0] = 45$ and $\Delta = 0.04$..

For completeness, Table 8.5 shows the equivalent measurements on the SI for a series of scores on each of the instruments. Figure 8.1 illustrates this equivalence for CHE. Table 8.5 shows that a score of 50 on ENG, for example, is a measurement of

Table 8.5 Equivalent measurements on the SI for the same score on each of the instruments

Score	ENG SI	LIT SI	MA1 SI	MA2 SI	HIM SI	CHE SI
20	15.3	26.0	18.8	27.6	18.7	15.3
30	16.8	26.4	28.0	34.4	22.9	26.9
40	21.7	26.8	33.6	40.3	27.5	33.5
50	31.8	27.6	37.6	46.2	33.8	38.5
60	44.7	34.2	42.0	52.9	43.4	44.5
70	59.4	49.6	48.6	61.5	57.9	54.6
80	74.8	71.9	59.8	73.0	78.8	72.3
90	90.1	101.4	78.0	88.5	107.6	101.1
95	98.9	119.5	91.0	99.1	125.8	121.0

Table 8.6 Pairwise latent correlations and descriptive statistics for each instrument from the total sample of 13617, anchored to principal components from the analysis of edited profiles at 0.85 and transformed according to $\beta_{SI} = \beta_i/0.04 + 45$

ρ_{ij}/r_{ij}	ENG	LIT	MA1	MA2	HIM	CHE
ENG		*	0.9723	1.0160	0.9680	0.9907
LIT	*		0.9836	0.9989	0.9484	1.0092
MA1	0.2987	0.2632		1.0592	1.0389	1.0192
MA2	0.4418	0.4096	0.9845		0.9187	1.0348
HIM	0.6759	0.7483	0.4881	0.6510		1.0338
CHE	0.4640	0.5167	0.8496	0.8698	0.7344	
N	10974	1463	4426	1594	2014	4973
Mean	43.25	54.68	49.61	57.05	47.17	47.86
St Dev	14.26	18.24	16.41	17.19	15.58	17.33
Skew	0.48	0.83	0.22	0.17	0.62	0.49

Note. *Frequency of less than 20 not shown. Mean of the pairwise latent correlations for homogeneous profiles: $\overline{\rho}_{ij} = 0.9994$. Some observed correlations are slightly greater or less than 1 due to random variation around 1. Mean of the pairwise latent correlations for measurements of all data: $\overline{r}_{ij}$ =0.5997

31.8 on the SI, while the same score on MA2 is a measurement of 46.2. For scores up to 70, the greatest measurement on the SI is for MA2, the advanced mathematics discipline. For scores greater than 70, HIM (modern history) has greater measurements on the SI. This order results from it being more difficult to obtain a very high score in HIM than in MA2 for the respectively very proficient students. It is not uncommon for it to be difficult to obtain very high scores in humanities disciplines, while it is much more common to obtain very high scores in the mathematics and science disciplines. The EVCs of Fig. 8.5 reflect the relative difficulty at the higher end of the proficiency continuum.

Table 8.6 shows two sets of latent correlations and the distribution properties of measurements on the SI for scores on each instrument. Above the diagonal, it shows the latent correlations corrected for errors of measurement for the analysis of the edited profiles, which results in homogeneous scores. It will be recalled that this analysis ensures that the equating functions are obtained from profiles that are effectively unidimensional. The average latent pairwise correlation of 0.9994 ensures that the profiles are homogeneous. Of the 5708 profiles that had at least two measurements after the profiles were edited, 24% had a standard deviation less than 5. Given the unit of the SI, $\Delta = 0.04$, which implies a variance of $\sigma^2 = 1/\Delta = 25$ for replicated measurements in the linear range between 15 and 86, where the majority of measurements are, it would be expected from a Gauss distribution that some 32% would have a standard deviation less than 5. Thus if anything, the profiles are slightly more homogeneous than under total randomness.

Below the diagonal, Table 8.6 shows the latent correlations between the instruments for measurements of all profiles of all persons. Clearly, with all profiles measured, the latent correlations are not homogeneous, and of course not close to 1. As expected, with a mean of 0.5997, they mirror the observed correlations between the raw scores shown in Table 8.3. Such correlations are reflected in non-homogeneous profiles. Of the 7334 profiles with two or more scores, 55.83% have a standard deviation greater than 5 which indicates that they deviate from their respective means by approximately five score points. For example, a profile with two scores and a standard deviation of 5, has measurements of 48.90 and 58.90. 22.08% have a standard deviation greater than 10, which indicates scores that essentially deviate 10 points in either direction from their respective means. For example, a profile with two measurements and a standard deviation of 10.00 has measurements of 69.30 and 89.30.

The profiles with a standard deviation greater than 10 are not characterized by their total scores. This seems very relevant in the system of selection, which is based primarily on the total score. Depending on which course of study the student is planning, those profiles with a marginal case for selection on the basis of their total score, would need to be considered individually. For example a relatively low score in the required discipline of English and relatively high scores in the mathematics and science disciplines, may not preclude a student marginal for selection on the basis of the total score, being selected for an engineering course. It is stressed that in formal university entry, at least four measurements, including one of ENG or LIT, are required to meet the eligibility requirements for university entry. Then the application is based on the mean of the four highest scaled measurements. In the illustrative example of this chapter, where the maximum number of measurements is only six, very few people have the minimum four and therefore the means and rankings from these illustrative data are not useful to study.

Recognizing that many profiles are not characterized by their total scores, Table 8.6 shows the proficiencies of all students on all instruments. The measurements on MA2 have the greatest mean (57.05), followed by LIT (54.68). It will be recalled that these are more advanced specialist disciplines in mathematics and English respectively, and therefore it is expected that their relative means will be the greatest. That they are, also confirms the success of the equating. It is noticeable, however, that the mean of LIT, which was 70 from the raw scores and the largest, is no longer the largest. ENG, the discipline taken by many students because English proficiency is required for university entry, has the smallest mean among this group of students.

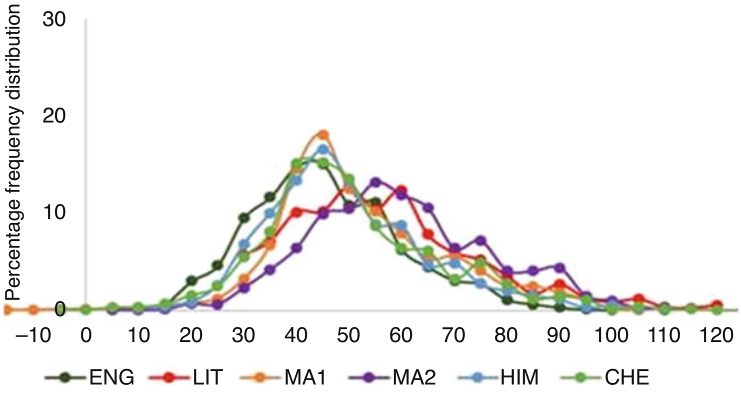


Fig. 8.6 Distributions of equated scores to a standard instrument in class intervals of 5 score points

Figure 8.6 shows the distributions of these measurements in class intervals of five score points. It reflects the distributional statistics in Table 8.6, with the standard deviations relatively homogenous. On the other hand, unlike the original scores, which show a negative skew, the measurements on the SI all show a positive skew. It is noticeable that only the distributions of the two mathematics disciplines, MA1 and MA2, have a skew value clearly less than 0.5, suggesting they are the only ones with clearly normal distributions, while the others, and in particular LIT, show deviation from normality. A final point to observe is that the distributions of the two specialist disciplines, LIT and MA2, show similar features, while the distributions of the other disciplines are also similar.

8.4 Summary and Discussion

This chapter began by observing two contrasting but interrelated aspects of the idea of measurement. First, that the historical motivation for physical measurement, the standardization of a unit and origin within a frame of reference, was fairness of transactions of goods for trade in everyday applications. The aim of such measurement was to ensure that comparisons among objects, for the relevant property, were invariant with respect to which specific instrument was employed. Second, that although in its elementary form, measurement is understood by school children, the mapping of the magnitude of a property of an object onto a number line partitioned into equal units, is a deep abstraction, and that the advanced application of measurement has evolved in conjunction with the remarkable advancement of the quantitative, natural sciences. Although many scientific measurements remain in the realm of scientific theory, many have also become mainstreamed and are applied in everyday applications. Measurement of temperature is an example, where although integral to the theory of thermodynamics, thermometers for measuring human temperatures in the case of potential illness have become indispensable.

The success of quantitative laws in explaining phenomena in the natural sciences provides an aspiration for quantitative laws in the social sciences. The coevolution of measurement and quantitative laws of the natural sciences implies that any quantitative laws of the social sciences will evolve in conjunction with social measurement, where the concept of measurement transcends the natural and social sciences. However, there are frames of reference in the social sciences where the historical motivation for measurement in the natural sciences prevails – that of fairness of transactions. By analogy to the use of rigorous measurement of temperature in routine applications, measurement in the social sciences may require measurement which is just as rigorous as is the measurement of variables that will produce scientific laws. This chapter is concerned with such an example.

The frame of reference of this chapter is the competitive selection of students into universities in Western Australia based on their assessments in a range of disciplines in their Year 12 studies. The assessments are not of the form of intelligence or aptitude tests, but based on explicit syllabuses that the students have been taught. The students may choose to pursue their university studies in a wide array of courses based on their Year 12 studies in a similarly wide array of disciplines. In general, satisfactory performance in the discipline of English is required for selection into all university courses. Many courses, such as law, psychology, and economics, have no prerequisite studies, while some courses such as engineering, natural sciences, and mathematics, may require studies in those disciplines, but permit electives which may vary among students. The selection is based primarily on the mean of four performances, from the disciplines they have studied, which give the greatest mean. The implication of these features is that, in the interest of fairness, the selection of students for various university courses needs to be *invariant* with respect to which disciplines they studied. This requires rigorous measurement, with the same unit and origin, of each discipline.

It is recognized that, because different students may have studied different disciplines, their summary scores do not reflect a degree of proficiency on the same substantive, content variable. Instead, the variable is an abstracted index variable of causal variables in which the proficiency of a student in a discipline governs their performance on the relevant assessment instrument. The variable is inferred to be one of a capacity to succeed and benefit from university studies based on previous studies. The example is described in some detail with the expectation that it has properties that can be transferred to other social sciences cases where rigorous measurement is required.

In part because of the advancement of the natural sciences with the coevolution of quantitative laws and measurement, social scientists have studied and attempted to define measurement generally from the perspective of advancing quantitative laws in the social sciences. One of these is the work of Rasch, termed in this chapter, *Rasch measurement theory*. The principle on which Rasch's theory is based is the requirement of invariant comparisons of objects with respect to instruments (and vice versa) within a specified frame of reference. The consequences of this definition have been shown to lead to quantitative relationships which are entirely compatible with other definitions of measurement and with the laws of the natural sciences.

Rasch's formulation has at least three distinctive elements compared to other definitions of measurement. First, it is relevant in both probabilistic and deterministic frameworks. The probabilistic contexts immediately provide evidence of statistical variation that may or may not be random. Second, because Rasch formulated his requirement of invariant comparisons mathematically, rather than merely descriptively, further derivations of the model for measurement are possible. Third, the requirement of invariant comparisons is relevant for quantitative laws in general, and not only measurement; therefore, it seems a more fundamental basis for understanding measurement than simply describing measurement. All three features of Rasch measurement theory are applied explicitly in the example of this chapter.

Rasch's probabilistic formulation is used in this chapter. It specifies the probability that a person will obtain a score on the instrument as a function of the person's hypothesized, scalar, proficiency and the threshold parameters of the instrument which reflect its relative difficulty (origin) and their tendency to spread and skew the responses. The thresholds are points of equal probability of two adjacent scores.

Then, given estimates of its parameters, the estimate of every person's proficiency on each instrument is transformed to an expected value on a SI (standard instrument) that is termed a *measurement*. The procedure was introduced and summarized in Fig. 8.1. The distribution of the SI has been derived from Rasch's original formulation and is unlikely to have been formulated in any other way. The origin and unit of the SI are as explicit as they are in measuring instruments in the natural sciences, and are chosen for convenience. In particular, over a substantial and defined range of the instrument, the difference between two successive measurements is the unit. In addition, in this range, the distribution of inferred replicated measurements of each person to each instrument is a discrete analogue of the continuous Gauss distribution in which the variance on the SI is the inverse of the unit. It was recalled that the Gauss distribution was derived to characterize random variation of replicated measurements of the same object with the same instrument. Finally, the region in which the relationship between the measurement on the SI and student proficiency is not linear is near scores of 0 and 100 where the limits of the instrument interfere with the random variation from replications. This is a region in which the Rasch distribution of the SI is applicable, but in which the Gauss distribution is not.

The origin and unit of the SI in the example were chosen to minimize the number of measurements in the region in which the proficiency and the expected value of the SI are not linear. The advantage of this choice is that the variance of the SI for any person reflects no more than random variation. Therefore, if the observed variance of the SI scores on a profile is greater than random variation, the total score does not summarize the profile. In this case, not only should the magnitude of the mean measurement of a profile be considered for selection, but the profile should also be studied for evidence of specific capabilities that might be relevant for the choice of further studies.

The motivation for transforming the scores of each instrument to a measurement on the SI is that of invariance of comparisons – that comparisons between students for competitive selection is invariant with respect to the subset of disciplines that they have studied from a wider set of relevant disciplines. This motivation, rather than that of advancing quantitative laws in the social sciences, is identical to the original motivation of much of measurement of physical variables. Importantly, although the motivation for invariant comparisons in this case is clearly that of fairness of selection, it is the same motivation that led to Rasch's measurement theory of invariance and which is relevant for understanding and constructing measurements which can lead to quantitative laws in the social sciences.

It was also indicated that, because the frame of reference was complex, the example can be taken as illustrative and that the approach taken to equating can be generalized to other frames of reference. One of these is person-centred outcomes in health assessment. As is evident throughout the example, the concern is with personal profiles and selection of individuals; therefore the example can be considered to be person-centred in its concern, an approach exemplified in Cano, Pendrill, Melin, and Fisher [14]. In the case of the assessment of an individual on multiple instruments in the health outcomes area, the principles, first that a summary score characterizes a higher order variable which is primarily an index variable, and second, that there will be individuals whose profiles are summarized by the total score and others which are not, is readily applicable.

In summary, it is stressed that, analogous to the Gauss distribution, the Rasch distribution of the SI employed in this chapter was derived from theoretical considerations, and not to describe any particular data set. Just as the Gauss distribution sets up a criterion that the variance of real or inferred replications is no more than random, and therefore that the mean can be used as a summary measure for the replications, the Rasch distribution sets up the criterion that the distribution of inferred replications from an instrument is no more than random, that the total score is sufficient to characterize the profile, and that the mean can be used to summarize the profile. By implication, comparisons which are invariant with respect to the instruments across profiles can be made. In short, it is a criterion for measurement. In conjunction with stressing that the Rasch distribution is derived as a criterion for measurement and not to describe any data set, it is stressed that the observed scores from instruments can only be transformed successfully to measurements if the data themselves permit such a transformation. To ensure such a possibility, extensive substantive empirical and theoretical work and understanding is required.

Acknowledgments The School Curriculum and Standards Authority, Department of Education, Western Australia, gave permission for the use of the data analyzed in the example. The research was supported in part by grants from the Australian Research Council.

References

1. K. Alder, *The Measure of All Things: The Seven-Year Odyssey and Hidden Error That Transformed the World* (Free Press, 2002)
2. E.B. Andersen, Sufficient statistics and latent trait models. Psychometrika **42**, 69–81 (1977)
3. D. Andrich, A rating formulation for ordered response categories. Psychometrika **43**(4), 561–574 (1978)

4. D. Andrich, A structure of index and causal variables. Rasch Measur. Trans. **28**(3), 1475–1477 (2014)
5. D. Andrich, The problem with the step metaphor for polytomous models for ordinal assessments. Educ. Meas. Issues Pract. **34**(2), 8–14 (2015)
6. D. Andrich, Chapter 7: Perceived health and adaptation in chronic disease: Stakes and future challenge, in *Advances in Social Measurement: A Rasch Measurement Theory*, ed. by F. Guillemin, A. Leplège, S. Briançon, E. Spitz, J. Coste, (CRCS Press/Taylor and Francis, 2018), pp. 66–91
7. D. Andrich, G. Luo, Conditional estimation in the Rasch model for ordered response categories using principal components. J. Appl. Meas. **4**, 205–221 (2003)
8. D. Andrich, I. Marais, Person proficiency estimates in the dichotomous Rasch model when random guessing is removed from difficulty estimates of multiple choice items. Appl. Psychol. Meas. **38**(6), 432–449 (2014)
9. D. Andrich, P. Pedler, Modelling ordinal assessments: Fit is not sufficient. Commun. Stat. **48**(12), 2932–2947 (2019a)
10. D. Andrich, P. Pedler, A law of ordinal random error: The Rasch measurement model and random error distributions of ordinal assessments. Measurement **131**, 771–781 (2019b)
11. D. Andrich, I. Marais, S. Humphry, Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. J. Educ. Behav. Stat. **37**(9), 417–442 (2012)
12. D. Andrich, B.S. Sheridan, G. Luo, *RUMM2030Plus: Rasch Unidimensional Models for Measurement* (RUMM Laboratory, Perth, 2020)
13. N.R. Campbell, *Physics: The Elements* (Cambridge University Press, 1920)
14. S.J. Cano, L.R. Pendrill, J. Melin, W.P. Fisher Jr., Towards consensus measurement standards for patient-centered outcomes. Measurement **141**, 62–69 (2019)
15. M. De Podesta, Absolute zero, in *Nothing*, ed. by J. Web, (New Scientist, 2013), pp. 164–173
16. O.D. Duncan, Rasch measurement in survey research: Further examples and discussion, in *Surveying Subjective Phenomena*, ed. by C. F. Turner, E. Martin, vol. 2, (Russell Sage Foundation, 1984), pp. 367–403
17. C. Eisenhart, Law of error I: Development of the concept, in *Encyclopedia of Statistical Sciences*, ed. by S. Kotz, N. L. Johnson, vol. 4, (Wiley, 1983a), pp. 530–547
18. C. Eisenhart, Law of error II: Development of the concept, in *Encyclopedia of Statistical Sciences*, ed. by S. Kotz, N. L. Johnson, vol. 4, (Wiley, 1983b), pp. 547–562
19. L. Finkelstein, Well-defined measurement – an analysis of challenges. Measurement **42**(9), 1270–1277 (2009)
20. W.P. Fisher Jr., A.J. Stenner, Theory-based metrological traceability in education: A reading measurement network. Measurement **92**, 489–496 (2016)
21. H. Gulliksen, *Theory of Mental Scales* (Wiley, 1950)
22. S. Humphry, D. Andrich, Understanding the unit implicit in the Rasch model. J. Appl. Meas. **9**, 249–264 (2008)
23. Bureau Internationale des Poids et Mesures: Joint Committee for Guides in Metrology (JCGM/WG 1). (2008). Evaluation of measurement data--Guide to the expression of uncertainty in measurement. Sevres, France: International Bureau of Weights and Measures--BIPM. www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf.
24. M.J. Kolen, R.L. Brennan, *Test Equating, Scaling, and Linking: Methods and Practices*, 2nd edn. (Springer, 2004)
25. D.H. Krantz, R.D. Luce, P. Suppes, A. Tversky, *Foundations of Measurement*, vol 1 (Academic, 1971)
26. T.S. Kuhn, The function of measurement in modern physical science. Isis **52**, 161–190 (1961)
27. R.D. Luce, J.W. Tukey, Simultaneous conjoint measurement: A new type of fundamental measurement. J. Math. Psychol. **1**, 1–27 (1964)
28. G. Luo, D. Andrich, Estimating parameters in the Rasch model in the presence of null categories. J. Appl. Meas. **6**(2), 128–146 (2005)

29. I. Marais, D. Andrich, Formalising dimension and response violations of local independence in the unidimensional Rasch model. J. Appl. Meas. **9**(3), 200–215 (2008)
30. L. Mari, Epistemology of measurement. Measurement **34**, 17–30 (2003)
31. G.N. Masters, A Rasch model for partial credit scoring. Psychometrika **47**, 149–174 (1982)
32. G.N. Masters, B.D. Wright, The partial credit model, in *Handbook of Item Response Theory*, ed. by W. J. van der Linden, R. K. Hambleton, (Springer, 1997), pp. 101–121
33. J. Michell, Measurement: A beginner's guide. J. Appl. Meas. **4**, 298–308 (2003)
34. R. Ostini, M.L. Nering, *Polytomous Item Response Models*, Sage University Paper Series on Quantitative Applications in the Social Sciences (07-144) (Sage Publications, 2006)
35. J.O. Ramsay, in *Review of Foundations of Measurement, Vol. I*, ed. by D. H. Krantz, R. D. Luce, P. Suppes, A. Tverskey, vol. 40, (Psychometrika, 1975), pp. 257–262
36. G. Rasch, *Probabilistic Models for some Intelligence and Attainment Tests* (Danish Institute for Educational Research, Copenhagen, 1960). Expanded edition (1980) with foreword and afterword by B. D. Wright. The University of Chicago Press. Reprinted (1993): MESA Press
37. G. Rasch, On general laws and the meaning of measurement in psychology, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV*, ed. by J. Neyman, (University of California Press, 1961), pp. 321–334
38. G. Rasch, On specific objectivity: An attempt at formalising the request for generality and validity of scientific statements. Dan. Yearb. Philos. **14**, 58–94 (1977)
39. A.J. Stenner, D. Burdick, M.H. Stone, Formative and reflective models: Can a Rasch analysis tell the difference? Rasch Measur. Trans. **22**, 1059–1060 (2008)
40. A.J. Stenner, W.P. Fisher, M.H. Stone, D.S. Burdick, Causal Rasch models. Front. Psychol. **4**, 536–557 (2013)
41. S.M. Stigler, *The History of Statistics: The Measurement of Uncertainty before 1900* (The Belknap Press of Harvard University Press, 1986)
42. D. Surla, Application of the Rasch Model of Modern Test Theory to Equate Multiple Tests Using Their Total Scores. Unpublished PhD dissertation, The Univesity of Western Australia (2020)
43. L. Tesio, Items and variables, thinner and thicker variables: Gradients, not dichotomies. Rasch Meas. Trans. **28**(3), 1477–1479 (2014)
44. L.L. Thurstone, *The Measurement of Values* (University of Chicago Press, 1959)
45. J. Tognolini, D. Andrich, Analysis of profiles of students applying for entrance to universities. Appl. Meas. Educ. **9**(4), 323–353 (1996)
46. M. Wilson, G.N. Masters, The partial credit model and null categories. Psychometrika **58**, 87–99 (1993)
47. B.D. Wright, A history of social science measurement. Educ. Meas. Issues Pract. **16**(4), 33–45 (1997)

Chapter 9
Addressing Traceability in Social Measurement Establishing a Common Metric for Dependence

Thomas Salzberger

Abstract Measurement in the social sciences is typically characterized by a multitude of instruments that are assumed to measure the same concept but lack comparability. Underdeveloped conceptual theories that fail to expose a measurement mechanism are one reason for the incommensurable measurements. Without such a mechanism measurements cannot be linked to a fundamental reference as required by metrological traceability. However, traditional metrological concepts can be extended by allowing for direct links between different instruments, so-called crosswalks. In this regard, Rasch Measurement Theory proves particularly useful as it facilitates a co-calibration of different instruments onto a common metric. The example of the measurement of nicotine dependence through self-report instruments serves as a showcase of the problems in social measurement and how they can be overcome contributing to metrological traceability in the social sciences.

Keywords Instrument equating · Nicotine dependence · Scale interpretation · Crosswalks

9.1 Introduction

Proper measurement is a sine qua non for quantitative research in any field of scientific enquiry [14]. The social sciences, encompassing a range of academic disciplines, such as health, education, psychology, or business, are no exception. As diverse as these fields of science are, when it comes to measurement they share one common characteristic: many properties of interest reside in, or are attributes of, human beings. In education such a concept of interest might be some sort of proficiency, in business research it could be corporate image perceived by consumers, in health science quality of life, while self-esteem would be an example from psychology.

T. Salzberger (✉)
Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien, Wien, Austria
e-mail: Thomas.Salzberger@wu.ac.at

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3_9

In some instances, third persons (e.g., teachers, parents, clinicians, etc.) might act as raters and thereby produce the data. However, the majority of measurements in the social sciences, or social measurements for short, are based on self-report data, where items are presented to respondents (students, patients, consumers, etc.), who select the most appropriate answer option among a set of ordered response categories, mostly in the order of two to seven options. Typically, a number of items form a set, referred to as a scale or a measurement instrument, supposed to indicate values expressive of the same variable of interest, or construct. While the responses to items are observable, or manifest, the concept of interest is represented by an unobservable, hence latent variable. The inference of measured constructs never observed directly takes place whenever instruments are read, and so applies in all sciences.

Measurement is concerned with the inference of a magnitude of the latent variable from observable responses to manifest items. To enable this inference, the operation of measurement requires a measurement model that links observed responses to a latent variable. In social measurement, such models are also called psychometric models. Psychometric models are purely formal, statistical and “technical”. Their purpose is not to describe data. Rather, they need to embody requirements of quantitative inference. Thus, not every arbitrary model may serve the purpose of quantification. For measurement to be valid, the items also need to adequately represent the concept of interest. The assessment of validity is concerned with the question whether an instrument actually measures what it is supposed to measure. Consequently, to tackle the problem of validity, we need a conceptual theory that explains the concept of interest. Qualitative as well as quantitative evidence will then show whether the instrument indeed measures the variable of interest.

As social sciences mostly subscribe to a realist view [50, 57], their concepts of interest are, in the end, considered objectively given. They do not emerge from, or in any way depend on, our measurements. Rather, their existence gives rise to their measurement. Therefore, the same concept of interest should be measurable by different instruments, provided the concept exists as a quantitative construct. A range of instruments measuring the same concept can be beneficial to science, as instruments vary in terms of their range of operation, applicability under different circumstances or acceptability by respondent groups. From this it follows that if different instruments are supposed to measure the same concept, they need to refer to, or rather be inferred from, the same conceptual theory.

In practice, the situation is far from being straightforward. Many concepts lack universal agreement on exact definitions. Different instruments claiming to address the same concept may therefore measure somewhat different concepts. Conversely, instruments that do measure the same concept may bear different labels suggesting they are measuring different concepts (for example, to lend a unique selling point to a new scale). To complicate matters further, conceptual theories are sometimes rather vaguely defined. In some instances, they represent a post-hoc summary of the content of the items rather than a guiding principle for generating items. This problem can be illustrated by the difference between content validity and face validity[1] in the sense it is nowadays in use. Content validity means that all aspects present in the conceptual theory are adequately captured by items in the instrument. In contrast, face validity only asks whether what is included in the instrument "on the face of it" seems to be reasonable given the stated concept of interest.

The intricacies of concept definition are paralleled by a multiplicity of measurement models with some models failing to incorporate fundamental principles of quantification and, thus, to provide generalizable measurements that would be meaningful and interpretable beyond a particular study or even within one study. The conceptual and modelling problems have led to a very unsatisfactory state of affairs in social measurement. Measurement is not always properly justified, and measurements of the same concept lack comparability. The reason for the latter is that most measurements, even when theoretically and empirically supported, are expressed in a metric that is unique to the instrument they are based on. The lack of comparability prevents the social sciences from capitalizing on multiple instruments, which are applicable under various conditions, measuring the same concept. Even worse, incomparable measurements are a major obstacle to the synthesis of research findings.

Currently, only very few research initiatives in the social sciences have addressed this problem, which deserves and requires much broader attention. Lifting social measurement to a higher level by raising both its conceptual as well as psychometric rigor, and providing comparable measurements by establishing common metrics ought to be the core objective of social measurement science in the twenty-first century. This agenda is also in the interest of the recent rapprochement of social and natural sciences in terms of measurement. Social sciences have realized that fundamental principles of metrology, the science of measurement, specifically metrological traceability and measurement uncertainty, apply to all sciences. Thus, these principles need to be understood properly and addressed in the context of social measurement.

In the following, a brief history of measurement in the social sciences sheds light on the roots of problems social measurement is currently facing. In terms of a measurement model, the Rasch model [67] lends itself as a theoretical framework that certainly has the potential to lead social sciences out of the impasse they have ended up in regarding measurement. Next, we discuss the challenges of metrology, notably metrological traceability and measurement uncertainty, and how the social sciences can address them. One way of addressing traceability is to link various instruments by equating and thereby establishing a common metric. The example of measurement of self-reported nicotine dependence illustrates this approach.

[1] Originally, face validity, which has not always been defined uniformly, mostly meant that respondents recognize the concept the items are supposed to measure [44]. As such, face validity is not necessarily desirable [16], for example in projective measurement. Today, face validity is typically assessed from the researcher's viewpoint, and is widely seen as an indication of but not a substitute for content validity.

9.2 Recourse to Literature

9.2.1 The Significance of Scientific Measurement

Scientific measurement is a clearly defined concept: it is the inference of a magnitude, an amount of a property supposed to be quantitative [59]. The natural sciences have made unprecedented progress based on quantitative theories and models [60]. Without measurement none of the theories could have been tested empirically. Today, a world without measurement has become inconceivable. In a sense, measurement is both a very simple and a very complex matter at the same time. It is complex inasmuch as quantity is probably one of the boldest propositions of an object's attribute one can think of. However, once measurement has become a successful technology with instruments available readily, it summaries the variable in a very simple manner: by one number.

9.2.2 Measurement in the Social Sciences

9.2.2.1 The Challenge of Social Measurement

The social sciences have tried to follow the natural sciences in their footsteps as becoming quantitative as this seemed to be the sine qua none of science in the twentieth century [60]. However, the path towards quantitative science taken in the social sciences was anything but straightforward. How should psychological concepts, or constructs, be measured, and how should there being quantitative be demonstrated? Michell [59] disentangled the problems of inferring measurement by suggesting two tasks to be fulfilled: the *scientific task of measurement* and the *instrumental task*. The scientific task is concerned with whether the variable to be assessed is actually quantitative. The instrumental task pertains to the procedures of estimating the magnitude, the amount of the property to be measured. The scientific task is more fundamental providing the basis for technologies performing the instrumental task.

While these tasks arise regardless of the scientific discipline, the challenges involved vary, and certainly very different procedures are called for depending on the field of research. As an illustration from the physical sciences, the measurement of temperature [22] has to explicate what exactly temperature is and demonstrate that it is quantitative. Evidence of the latter allows for a range of instruments to be designed for its measurement. Today, physical measurement seems to be largely a matter of technology and for the most part it arguably is. But it should not be overlooked that the understanding of many physical attributes often took centuries with temperature being a good case in point. On the other hand, fundamental underpinnings of measurement are continuously advanced as demonstrated by the redefinition of fundamental units in physics [19]. Likewise, new problems and applications, such as self-driving cars, occur that require new measurement technologies [41].

9.2.2.2 The Path Toward Quantitative Social Sciences

At the beginning of the twentieth century, demonstrating the quantitative nature of psychological concepts such as attitudes, perceptions, feelings, or emotions—that is, demonstrating that the fundamental requirement of any measurement to be meaningful can be satisfied—appeared to be too enormous a challenge [60]. Though the theoretical terms predominating today were proposed much later, in retrospect it is plain that the social sciences struggled from the very beginning with addressing the scientific task of measurement. The path chosen in the social sciences as a way of resolving the difficulties encountered in meaningful quantification have appeared to be very effective. Stevens' [79] suggestion that measurement is the assignment of a numeral to an observation according to a rule has been embraced widely. It essentially means that a numerical score is assigned to observed behaviors, such as the response to an item by ticking a box, and that the score is interpreted as a measure of some quantity. Today, social scientists, for the most part, subscribe to this definition without hesitation or reservation, and arguably with little awareness of its limitations. Measurement has become ubiquitous in the social sciences based on Stevens' definition.

9.2.2.3 Measurement by Assigning Numerals

But are all these measurements justifiable? Can measurement be achieved by simply assigning numerals and claiming they are measures? One could argue that even the natural sciences measure by assigning values. For example, an analog bathroom scale comes with a kilogram scale that assigns values to the deflection of the pointer – values which we then use as measurements. However, constructing a bathroom scale is merely an instrumental task that builds upon established evidence of mass being quantitative and links the pointer's deflection to a known metric of mass. In social measurement, the equivalent of such a bathroom scale would be a device built from scratch that somehow produces a deflection of the pointer which "defines" measurement by assigning numerals. Stevens' definition essentially invokes the instrumental task and implies that the scientific task has been addressed if the instrumental task is carried out properly ("according to a rule").

Stevens, though, fails to provide stringent guidelines as to how rules of assignment ought to be designed to yield measures of a quantity that are on a linear interval scale. Rather, he proposed different scale levels (nominal, ordinal, interval, ratio, and log-interval) that are supposed to reflect the true properties of what is to be assessed. The scheme is internally consistent but it is descriptive and not instructional. Without a clear rationale of whether numerals are interval-scaled or merely ordered, it is of limited use. Stevens arguably would be very skeptical of many social measurements explicitly or implicitly invoking his definition. But as a matter of fact, measurement by mere numerical assignments has opened the door to questionable practices in the social sciences. It definitely has contributed to a divergence of the social from the natural sciences in terms of a fundamental concept of science.

The following considerations apply to measurement of constructs in the social sciences inferred from data that are based on responses of subjects to stimuli, referred to as items, which collectively represent the construct to be measured. The items work as a set, a scale, and are proposed to be a measurement instrument. Instruments are included in questionnaires and surveys; they are administered on paper or online; they are self-administered or implemented by an interviewer. In any case, the observed responses are coded numerically. Thus, the administration of such an instrument necessarily provides numbers such as item scores and total scores across items. Traditionally, such scores are considered measures. However, observed scores can only be the input to measurement. Measurements are inferred from observed scores – provided they meaningfully represent a quantitative property. Automatically and uncritically equating observed scores with measurements means neglecting the scientific task of measurement.

9.2.2.4 The Quest for a Measurement Model for the Social Sciences

Addressing the scientific task of measurement in the social sciences is a matter of defining an appropriate measurement model that links inferred measurements and observed scores in a way as to require the property to be quantitative. The model has to specify restrictions imposed on observed responses that ensure that the inferred measures are magnitudes of a quantitative property. While the model is statistical in nature, the goal is not to account for the data in the best possible way but to challenge the data with respect to the requirements of quantity.

Since linking magnitudes and observed scores is a purely formal, statistical problem, the model is not specific to some particular variable. Rather it pertains to a specific type of data as the input. The data collection can be seen as a consequence of the instrumental task. The scientific task and the instrumental task are therefore intertwined in the social sciences at least initially. The fact that the measurement model is blind to the content of the property to be assessed implies that social measurement is both qualitative (are we measuring the right quality) and quantitative (are the measures truly quantitative). Thus, social measurement necessarily calls for mixed methods combining qualitative and quantitative research. Currently, the mixed methods complement one another. In the future, these methods should be more integrated as will be pointed out later.

It should also be noted that any proper measurement embodies a theory stating that the proposed property is quantitative (it *can be* measured; scientific task) and that the suggested instrument is capable of measurement (the property *is* measured; instrumental task). Scientific theories can, of course, be right or wrong. Measurement is no exception. Also, we can never prove that a theory is correct. Rather, failing to reject a theory increases our confidence in the theory and measurement. When we say measurement is valid, it is always a tentative statement confined to the context (population of subjects, method of data collection, etc.) and subject to further investigation.

9.2.3 The Rasch Model for Measurement in the Social Sciences

9.2.3.1 From Population-Based Score Statistics to Invariant Measurement

In terms of measurement models, social science has long relied on classical test theory (CTT; [52]), which essentially states that an observed score is composed of a true score and an error score. As such, CTT cannot be disproved, as explaining one observable score by two unobservable ones is tautological. CTT heavily focuses on group-based statistics, such as correlations of item scores or reliability defined as the ratio of true score variance and total variance. Reliability can be estimated, with some assumptions, from repeated measurements. In the end, though, CTT is a theory of error that presumes measurement. Later, in the twentieth century, a range of innovative measurement models, especially item response theory (IRT; [29]), were developed. IRT comprises probabilistic models that link the probability of a person choosing a particular response to an item (in the case of dichotomous response scales agree or disagree, for instance) to item and person parameters by means of a logistic function. The function relates theoretically unbound quantitative measures of the item and the person to a response probability between 0 and 1. IRT seems to have shifted the focus from groups, or samples, to the individual measurement.

However, most IRT models include a parameter describing item discrimination that can only be estimated under the assumption of a person distribution (typically a normal distribution). Consequently, the estimation of an item discrimination parameter makes IRT models multivariate group models. They are statistical models that "have little credibility as scientific models" [27, p. 23]. One type of model – the Rasch model [67] – named after the Danish statistician Georg Rasch, stands out as it requires item discrimination to be equal across items. Item discrimination is implicitly set to unity and not empirically estimated. As a result, item and person parameters can be separated; that is, the estimation of the item parameters does not depend on the estimation of the person parameters and vice versa.

9.2.3.2 Accounting for Measurement Requirements

The property of parameter invariance is the defining characteristic of a Rasch model for measurement. Rasch himself referred to it as specific objectivity [68]. However, requiring item discrimination to be equal across items is not just a convenience, a statistical trick, in terms of parameter estimation. Rather, it ensures that the ordering of all item response categories in terms of their probability is the same across the entire continuum of the latent variable and, hence, for all respondents. A change in the order would imply that one item is relatively harder to endorse than another item for one person but relatively easier to endorse for another person. The property of equal item discrimination in the Rasch model therefore reflects a fundamental requirement of meaningful measurement. Models estimating item discrimination, or to be specific, differences in item discrimination, mask counterintuitive and implausible relationships between item and person properties which render the observed responses unsuitable for inferring measurements reminiscent of Ragosa's [66, p. 193] conclusion that "models for relations among variables [are to be seen as] statistical models without a substantive soul." Of course, the Rasch model is also a statistical model, but it retains substantive interpretation of all parameters combining statistical and scientific elements in psychometrics [87].

9.2.3.3 Rasch Measurement Theory as a Framework for Quantification in the Social Sciences

The unique properties of the Rasch model ultimately led to suggesting the term Rasch measurement theory (RMT; [7]). RMT is a formal theory of inferring measures, which are linear and interval-scaled, from observed scores, which code manifest observations. Proposing the label of a measurement theory can only be justified if the theory has properties distinguishing it from other theories in this regard. Therefore, it is worth briefly delineating the relationships of RMT on the one hand, and CTT and IRT on the other. Both RMT and CTT make use of the simple unweighted raw score across items. In CTT, the raw score is a priori considered a measure, the behavior of which is investigated by means of group-based statistics such as variances, correlations and factor loadings.

In contrast, in RMT, the raw score is a sufficient statistic for the person measures with its meaningfulness being subject to data fitting the model. In a sense, RMT provides the justification of the raw score CTT relies on. RMT and IRT are obviously related by virtue of sharing the same type of link function, namely the logistic function. However, the statistical resemblance blurs far-reaching differences in the substantive underpinnings of RMT versus IRT. As pointed out above, the group-dependency of the discrimination parameter estimation in IRT implies that neither parameter invariance nor raw score sufficiency hold as equal item discrimination and raw score sufficiency are mutually dependent properties of the Rasch model [34]. Non-Rasch IRT models therefore do not allow for a simple raw score to measure conversion. While this might be seen as merely a pragmatic flaw, the lack of a unique ordering of items seriously compromises the interpretation of measures.

In this context, it is once again important to stress the difference between assumptions and requirements. RMT is a theoretical framework that embodies scientific requirements in a statistical model. These requirements are subject to tests of fit of the data to the model, hence they are falsifiable. They are not untested assumptions. Specific objectivity as a property of the model does not mean that empirical item parameter estimates are necessarily the same in the Rasch model, not even up to random error, regardless of who the respondents are. Rather, objectivity is confined to a frame of reference, for which invariance holds – hence the expression specific rather than general objectivity. Within the frame of reference, though, item parameter estimates are, in principle, independent of the sample composition.

The significance of the properties of the Rasch model is further demonstrated by their relationship to independence axioms as the bedrock of quantification and a cornerstone of the scientific task of measurement. The weak respondent independence axiom states that if a person A has a higher probability of success (or endorsement) with one item compared to person B, then person A has a higher probability on all items. The weak item independence axiom requires that if an item X implies a higher probability of success (or endorsement) than item Y for one respondent, then this has to be true for all respondents. Together these independence axioms are summarized as the single cancellation condition [47] establishing order relationships. IRT models only comply with weak respondent independence but generally do not meet weak item independence. Even more important are higher order cancellation conditions as they substantiate metric scales by enforcing transitivity of differences. Double cancellation involves sets of three items and three persons and requires one implication to hold true based on two premises. Details can be found in Karabatsos [47] and Salzberger [71].

Notwithstanding the special properties of the Rasch model, it has to be noted that RMT cannot, and indeed should not, imply or ensure measurement where the data do not allow for the inference of measurements. The model is the benchmark, an ideal against which empirical data are checked. The fit of the data to the model determines whether the attempt at measurement can be deemed successful or not. As fit will never be perfect, fit should be interpreted in a relative fashion. Fit, or the degree of fit, determines our confidence in measurement. This situation mirrors scientific practice in general. A single study neither confirms nor disconfirms a theory once and for all. Rather, it strengthens or weakens our confidence in the theory. The same is true for measurement. RMT equips social science with an appropriate tool to improve measurement by strengthening our confidence in well-performing instruments and exposing the weaknesses of less well-functioning scales.

RMT lends itself as the most suitable theoretical framework for measurement in the social sciences. Nevertheless, the intertwining of the scientific task and the instrumental task of measurement carries the risk that social measurements will not transcend the instrument used. The application of the Rasch model to data based on one instrument may provide evidence that the hypothesis of measurement can tentatively be confirmed. However, under different circumstances different instruments measuring the same property may be suited. In fact, it is the rule rather than the exception that multiple instruments exist to measure some property. The same, of course, is true in the natural sciences, where for example, a plethora of instruments measure temperature. But all these instruments express measurements in the same unit of measurement despite very different mechanisms used in the instruments. Consequently, temperature measurements are comparable regardless of the instrument provided they are all calibrated properly. Today, such measurement systems are extremely rare in the social sciences. Dealing with measurements of unknown quality and incomparable metrics is a major obstacle to progress in the social sciences. While linking different instruments supposed to measure the same property is a major challenge in social measurement, there are different avenues to pursue this goal. One calls for a stronger connection between the qualitative theory of the construct and the quantitative property, another exploits the invariance property of the Rasch model.

9.3 Metrology in the Social Sciences

9.3.1 Metrological Traceability

Comparable measurement lies at the core of metrology, the science of measurement in any field of science, with the focus on metrological traceability [12]. Traceability means that the result of measurement can be related to a reference through an unbroken chain of calibrations. A reference can be a definition of a measurement unit, a procedure or some standard. Metrological traceability ensures that measurements of the same property can be expressed in the same, hence comparable, metric. The measurement of mass nicely illustrates the role of the reference. Recently, the kilogram has been redefined based on Planck's constant [80] linking the unit to a natural constant in the same way as the meter has been linked to the speed of light in vacuum [37]. This definition allows for the realization of the unit of mass through defined procedures. However, before 2019, the unit of mass was defined by an international prototype, a platinum alloy cylinder, stored in Paris [61]. This prototype was a standard in the sense that every measurement of mass had to be related to all the others through a chain of calibrations. Inevitably, every calibration (for example a national prototype against the international prototype; an intra-national prototype against the national, etc.) contributes to measurement uncertainty, which will be discussed later. The same is true for the new definition, even though it entails unique advantages, such as a stable reference (by contrast, the prototype itself has not remained completely stable over time) and realizations of fractions of a kilogram for different purposes of measurement.

Metrological traceability in the physical sciences can serve as a template for social measurement. It implies measurements based on various instruments are traceable to a common reference. To make this possible, we first need to develop a common reference to which different instruments can be related. One candidate for the common reference is the conceptual theory of the construct to be measured. It defines the variable and sets out its structure. However, today, conceptual theories are typically qualitative. At best, the theories suggest an order of items. As such, instruments cannot be quantitatively linked to conceptual theories. To overcome this problem, the conceptual theory needs to specify a quantitative link between a fundamental principle and properties of the measurement instrument. Specifically, that means revealing a measurement mechanism that explains item properties. Currently, examples of such advanced conceptual theories are still scarce [1, 35, 58, 76, 77, 78, 82].

A measurement mechanism in the social sciences corresponds to the most current definitions of units in the physical sciences, for example one meter as the unit of length defined as the distance light travels in a given interval of time in perfect vacuum. However, prior to this definition, an international prototype was used and referred to as a sort of "gold" standard of length. In a similar vein, the best available measurement instrument in social sciences could be seen as a gold standard [86]. Indeed, it is not uncommon in the social sciences to perceive one instrument as the undisputed benchmark of measurement. Thus, a reference would be established essentially by a consensus of scholars who agree on the gold standard similar to agreeing on a prototype of one kilo or one meter, the original definitions of the units of mass and length. Of course, the analogy is not perfect as the prototypes of one kilo and one meter demonstrably are magnitudes of quantitative properties. In the social sciences, the selection of one instrument as the gold standard is consensus-based. However, current psychometric models offer much stronger support for measurement instruments in the social sciences than ever before.

In terms of metrological traceability [12] in the social sciences, measurement mechanisms represent one possible realization of a fundamental reference to which various measurement instruments can be linked. As each instrument would be directly related to a reference that transcends all instruments and represents the common principle of all instruments, we might call this type of linking vertical linking. Different instruments would then be related indirectly via their linkages to the common principle. Alternatively, instruments can be linked to one another directly. Here, the instruments involved would be on the same level. Hence, we might refer to this as horizontal linking. It has to be noted that in the physical sciences no such differentiation (vertical versus horizontal) is made as traceability always implies an unbroken chain of calibrations against the fundamental reference as the root. From this it follows that addressing traceability in the social sciences by linking instruments with one another directly is an approach sui generis towards traceability. In the remainder of this chapter, the potential of linking instruments to another will be demonstrated.

9.3.2 Creating Measurement Systems

Providing for metrological traceability by linking different instruments measuring the same construct, that is the same measurand, means that measurements derived from different instruments are expressed in the same, hence comparable, metric. First, such a metric needs to be defined. The Rasch measurement model is particularly well suited to achieve this goal. The measurement instrument used in this regard corresponds to the gold standard mentioned earlier. For this task, it is therefore best to use the instrument with the most solid foundation in terms of its development.

Second, all instruments supposed to measure the same measurand need to be related to this metric. This can be done by co-calibrating the instruments with the gold standard instrument. In this regard, the invariance property of the Rasch model is particularly advantageous. As any set of items implies the same person measurement up to measurement error, scores on one instrument (one set of items) can be equated to scores on another instrument (another set of items) by aligning the respective estimates of person measurements. Particularly in the health sciences, the conversion of scores on one instrument to scores on another, mediated by an underlying common metric of the measurements, is referred to as a crosswalk [9, 85]. Metrological traceability among several instruments can then be established by a series of crosswalks. The instruments then form a measurement system.

Metrological traceability is by no means an end in itself. Rather, it is crucial for scientific synthesis of research findings. Crosswalks allow for linking existing findings based on different instruments that would otherwise be hard to compare. In case of a newly developed scale, crosswalks with existing instruments may also contribute to the acceptability and popularity of a new instrument by providing a smoother transition. Notwithstanding the practical potential of crosswalks, it has to be kept in mind that they do not shed light on the measurement mechanism. Crosswalks should be considered a transitory technology until conceptual theories are advanced enough to reveal the measurement mechanism.

9.3.3 Uncertainty

Apart from metrological traceability, a stated range of uncertainty [13] is a key feature of measurement. Every measurement, however carefully it may have been made, comes with some range of doubt as to its true value [10]. The range of uncertainty is a hallmark of quality in measurement. Uncertainty needs to be differentiated from error, which is the actual difference between the true value and the measured value, as it is referred to in CTT. Thus, CTT is effectively a theory of error, where error is the difference between the unknown true score and the observed score as the measurement. In every instance, the error score is a given but unknown value. If sources of error were known, the measured value could have been corrected. In contrast, uncertainty is concerned with our doubt as to how close the true value is to the measured value. Hence, uncertainty requires stating a range around the measured value in which the true value with some specified probability lies. The more sources of error impact on measurement, the larger the discrepancy between true values and measured values will be and, consequently, the wider the range of uncertainty will be. To know the range of measurement uncertainty is crucial as it determines whether the measurements are fit for purpose.

The sources of error in physical measurement are manifold [10], and they can be mapped onto social measurement, too. First, the measurement instrument can lead to error because of inadequate and/or outdated item wording or its unsuitable shape and appearance. This underlines not only the importance of careful instrument development but also stresses the necessity of continuously revisiting instruments over time. Second, the human beings whose properties we want to measure may show variation over time with respect to the manifestations of the measurand or the measurand itself. Changes in the measurand itself are particularly relevant when transient states rather than stable traits are to be measured. However, even if the true value of the measurand stays the same, inevitable fluctuations in the human mind, memory, and information processing in general imply variation in human responses to measurement instruments. Third, the measurement process may impact on the error emphasizing the significance of the context in which measurements are taken. For example, the administration of a questionnaire in a loud environment or under other inconvenient conditions arguably increases errors. Fourth, uncertainties in the calibration of the instrument itself contribute to error. The inference of measurements of person properties ultimately depends on accurate item calibrations.

The various sources of error show that careful instrument development and a suitable administration of the instrument are fundamental requirements of social measurement. That said, fluctuations in the respondents imply random variation in the responses of individuals that cannot be completely avoided. This variation is accounted for by the probabilistic element in psychometric models. The role of item calibration reveals an important insight into the mutual relationship of uncertainty in the item calibration and person calibration. Uncertainty of person measurements is reduced by item calibrations with low uncertainty. Conversely, item calibration uncertainty is a function of the sample size (the more subjects the better) and the sample composition. The closer the match of items and respondents in terms of their measurement values, in other words the better the targeting of the instrument, the more trustworthy item calibrations are. Proper targeting is therefore crucial for uncertainty. What is more, proper targeting also increases the power of tests of fit ensuring a more trustworthy assessment of the instrument’s suitability. The role of targeting also shows that uncertainty necessarily varies among individual respondents as the targeting of the instrument to the individual respondent depends on the respondent’s true value.

With all other factors considered given, uncertainty is only a function of how many measurements are taken. In the social sciences, this implies that with an increasing number of items, uncertainty decreases as error cancels out to some extent. In practice, this will only be correct up to a point, though. An extremely long instrument induces response burden, which may diminish concentration and increase rather than decrease error. In CTT, measurement uncertainty is derived from an estimate of reliability, which is defined as the true variance divided by the total observed variance, or by 1 minus the error variance over total variance. Combining reliability and the standard deviation of the person scores yields the standard error of measurement [84]. With perfect reliability, measurement uncertainty, that is the standard error of measurement, would be zero. With zero reliability, the standard error of measurement would equal the standard deviation of respondent scores. From this it follows that the standard error of measurement in CTT is both sample dependent and constant for all respondents irrespective of their true value.

By contrast, in the Rasch model the estimation of uncertainty with respect to the person measurement is based on the information the observed responses contain specifically for a person with a given total score across all items in the scale. For each item, information (INF) is a function of the probability of each response category. In the case of dichotomous items, information simplifies to the product of the two response probabilities (see Eq. 9.1; Fischer [33, p. 294]). In the polytomous case, information is a function of all item thresholds (see Stone [81], for a formula, or Dodd and Koch [26] and Muraki [62], for an alternative but equivalent formulation).

$$\mathrm{INF}\ (\beta_v, \delta_i) = P(\mathrm{a}_{vi} = 0) \cdot P(\mathrm{a}_{vi} = 1) \tag{9.1}$$

Information provided by each item adds up across items yielding total information ($\mathrm{INF}_{\mathrm{scale}}$) from the entire scale. The SEM then simply is the inverse of the square root of $\mathrm{INF}_{\mathrm{scale}}$ (Eq. 9.2).

$$\mathrm{SEM}\ (\beta_v) = \frac{1}{\sqrt{INF_{scale}}} \tag{9.2}$$

Considering the SEM approximately normally distributed implies that forming a confidence interval of the estimated person measurement ± SEM provides a range of measurement uncertainty with a confidence level of 68%. A probability of 95% can be achieved by multiplying SEM by 1.96 or 2 as an approximation. It should be noted that the symmetrical error distribution becomes rather implausible towards the extremes. Strictly speaking, no finite SEM can be estimated for extreme response patterns based on Eq. 9.2, as information would be zero. Therefore, the SEM for extreme scores is usually based on an extrapolation from SEMs for all other scores.

The way the Rasch model estimates the SEM essentially means that the range of uncertainty for each respondent not only takes the number of items into account but also their position relative to the respondent. As a consequence, information usually peaks close to the center of the scale, where the average distance to item thresholds is minimal, whereas uncertainty increases notably when approaching the extreme ends of the scale. The exact shape of the information curve depends on all thresholds, though.

9.4 Illustrative Example: Measurement of Self-Reported Nicotine Dependence

9.4.1 Purpose

As an illustration, the example of measuring nicotine dependence based on self-report instruments demonstrates the creation of a crosswalk, its usefulness, caveats and limitations.

9.4.2 Background and Literature

Nicotine dependence is one of several constructs that are important when it comes to understanding tobacco use [54]. The use of tobacco products, most notably cigarettes, still is the largest preventable cause of disease and premature death worldwide [46]. Tobacco-related health threats can best be prevented, or reduced, by cessation [83]. However, some tobacco users continue to use these products for various reasons. These users might benefit from less harmful products, also known as reduced-risk products, which also provide nicotine but otherwise reduce the amount of harmful chemicals [69]. Monitoring nicotine dependence in the transition phase may show how dependence is transferred from product to another and how it develops over time. As tobacco and/or other nicotine-containing products (TNPs) are subject to regulatory approval, authorities such as the Food and Drug Administration (FDA) in the USA also benefit from properly developed measurement instruments and traceable, comparable measurements of variables of interest.

The concept of nicotine dependence goes beyond physical addiction. It is a relatively complex construct consisting of behavioral and perceptive aspects. Apart from nicotine addiction as the key driver of dependence on TNPs, environmental and situational cues also have an impact [11]. The intricacy of measuring nicotine dependence is further exacerbated by the multiplicity of different TNPs available today. The first self-report instruments measuring nicotine dependence were developed for cigarettes, which historically were the predominant source of nicotine. Published in 1978, the eight-item Fagerström Tolerance Questionnaire (FTQ; [31]) was one of the first attempts in this regard. The FTQ was substantiated based on correlations with indicators of physical dependence (heart rate, body temperature) in experiments. Subsequently, the Fagerström Test for Nicotine Dependence (FTND; [42]) was developed as an improvement of the FTQ. The FTND is a six-item questionnaire related to smoking behavior. Due to its brevity and the meaningfulness of its content, the FTND has become one of the most widely used legacy measurement instruments in the field. Accordingly, it is often considered a gold standard for self-reported assessment in the field of nicotine dependence to the present day [65].

The context of measuring nicotine dependence through self-report instruments has changed, though, since the FTND was introduced. A range of alternative TNPs – such as smokeless tobacco or cigars or, more recently, electronic cigarettes – have become available [20, 21] and consumption of multiple TNPs in parallel can no longer be neglected [8]. In view of this development, the FTND was recently renamed as the Fagerström Test for Cigarette Dependence [32]. For alternative TNPs gaining in popularity, new self-report instruments for measuring dependence on these products have been suggested. (e.g., the Fagerström Test for Nicotine Dependence–Smokeless Tobacco questionnaire, FTND-ST [28], or the Penn-State Electronic Cigarette Dependence Index [36]). To complicate things further, product-specific instruments are not suitable for users of multiple TNPs concurrently. The multiplicity of instruments raises problems of comparability of measurements based on different instruments as each instrument has its own metric. From a metrology point of view, the incommensurability of measurements is a major shortcoming.

In order to overcome the limitations in comparable measurement of nicotine dependence, a new instrument, the ABOUT–Dependence self-report instrument,[2] was developed [23]. The self-reported instrument ought to capture the individual perspective of dependence on a range of TNPs as well as the concurrent use of

[2] The instrument is part of the ABOUT™ Toolbox initiative under which fit-for-purpose self-report instruments for assessing consumer responses to tobacco and nicotine products are developed [23, 24]

multiple TNPs. From the outset, the new scale was not aiming at the addition of another self-report instrument to the existing inventory of dependence scales but at providing the basis for comparable measurements of nicotine dependence irrespective of TNP usage patterns. The ABOUT–Dependence was to establish a metric of nicotine dependence to which existing legacy instrument can be referenced through "crosswalks".

9.4.3 Development of the ABOUT–Dependence Instrument

Based on a literature review, concept elicitation interviews with TNP users (n = 40) and input from experts [23, 24], a conceptual model of nicotine dependence comprising seven aspects of dependence experienced by TNP users (urgency to use upon waking up, compulsion to use, difficulty to cease using, need to use to function normally, automaticity of using, priority of using over social responsibilities, and self-awareness of dependence), thus comprising perceptual phenomena and self-reported behavioral aspects, was generated. The common theme behind these aspects is loss of control on the part of the product user. At this stage, a unidimensional structure of but no order relationships among the aspects of dependence were expected.

A first draft version of the instrument comprised 19 potential items that best represented the aspects of the concepts of interest was then analyzed by cognitive debriefing interviews (n = 40) to ensure proper understanding of instructions, items, and response options. While six items were identified as conceptually redundant, and one item appeared to be relevant only to some of the participants, all 19 items were subjected to a quantitative study and tested psychometrically to identify the best-working items. Three different response scales adapted to the item content were administered. Two items capturing the urgency and pervasiveness of product use over the past 7 days (see Table 9.1) were presented with a six-category scale (0–5 min, 6–15 min, 16–30 min, 31–60 min, more than 1–3 h, more than 3 h). Twelve items (eight in the final set) assessing the frequency of aspects of dependence over the past 7 days were administered with a five-category response scale (never, rarely, sometimes, most of the time, all the time). The remaining five items (two in the final set) referring to the intensity of current perceptions had a different five-category response scale (not at all, a little, moderately, very much, extremely). All studies were conducted in the United States.

9.4.4 Study Design, Data Sources and Sampling

The quantitative study, approved by the New England IRB (NEIRB#:120180022), consisted of a cross-sectional, two-wave, internet-based survey with purposive stratified sampling of adults legally authorized to purchase TNPs in the United

Table 9.1 Category frequencies of the ABOUT–Dependence items

Item (abbreviated)#	Complete sample (instrument calibration) n = 2434 frequency			Cigarette users (crosswalk sample) n = 250 frequency								
Extent-of-use	0–5’	6–15’	>3h	0–5’	>1 h–3 h	>3 h	0–5’	6–15’	16–30’	31–60’	>1 h–3 h	>3 h
[1] How long before going to sleep last product	281	411	487	547	447	261	28	50	58	53	42	19
[2] How soon after you woke up first product	314	498	470	411	339	402	37	63	58	44	25	23
Behavioral impact	Never	Rarely	Sometimes	Most of the time	All the time	Never	Rarely	Sometimes	Most of the time	All the time		
[3] Use more than you intended	357	621	855	416	185	19	67	106	35	23		
[4] Stop what you were doing to use product(s)	802	466	693	295	178	78	45	85	27	15		
[5] Use in a situation where you weren’t supposed to	1003	570	483	224	153	123	69	33	14	11		
[6] Sneak off to use product(s)	1150	424	509	210	141	139	35	47	21	8		
[7] Avoid an activity because you couldn’t use product(s)	1293	453	403	168	117	147	40	44	11	8		
Signs and symptoms	Not at all	A little	Moderately	Very much	Extremely	Not at all	A little	Moderately	Very much	Extremely		
[8] Strong desire to use product(s)	127	287	945	742	333	3	23	106	83	35		
[9] Difficult . . . to completely quit product(s)	313	358	496	572	695	9	33	52	64	92		
	Never	Rarely	Sometimes	Most of the time	All the time	Never	Rarely	Sometimes	Most of the time	All the time		
	315	384	924	526	285	14	41	111	52	32		

(continued)

Table 9.1 (continued)

Item (abbreviated)$^{\#}$	Complete sample (instrument calibration) n = 2434 frequency					Cigarette users (crosswalk sample) n = 250 frequency				
[10] Feel that you "HAD to have one"?										
	Not at all	A little	Moderately	Very much	Extremely	Not at all	A little	Moderately	Very much	Extremely
[11] Need product(s) to function "normally"	444	412	699	540	339	37	50	72	64	27
	Never	Rarely	Sometimes	Most of the time	All the time	Never	Rarely	Sometimes	Most of the time	All the time
[12] Hard to control the need or urge to use	531	510	740	421	232	39	52	94	47	18

States. Participants were identified via the proprietary GfK (Growth from Knowledge) consumer online panel KnowledgePanel®, which is representative of the US population. To ensure adequate coverage of the intended target population, quotas ensured that exclusive users of one TNP ($n = 1181$) and users of multiple TNPs (polyusers; $n = 1253$) were roughly equally represented implying a total sample size of 2434 respondents. Among exclusive users, an approximately equal number of users of cigarettes ($n = 250$), smokeless tobacco ($n = 250$), e-cigarettes ($n = 252$), and cigars/cigarillos ($n = 250$) were included, while the remaining 179 participants were exclusive users of pipes, waterpipes, or nicotine replacement therapy (NRT) products. Additional quotas on age, sex, and education were applied. Wave 2, which allowed assessment of stability over time (test–retest reliability), included 678 poly users and 743 exclusive users. Table 9.2 provides descriptive summary statistics of the sample of the quantitative study.

9.4.5 Psychometric Methods

The psychometric analysis was based on the unrestricted Rasch model for polytomous responses (partial credit model; [2, 4, 55]; see Eq. 9.3) using the computer software RUMM2030 [6], which applies the pairwise estimation algorithm [91] for item calibration and weighted maximum-likelihood estimation [51] for person calibration.

$$\mathrm{P}\left(\mathrm{a}_{vi} = \mathrm{x} \mid \beta_v, \tau_{ij}, j = 1 \ldots m, 0 < x \leq m\right) = \frac{e^{\left(\sum_{j=1}^{x} - \tau_{ij}\right) + x \cdot (\beta_v - \delta_i)}}{\gamma} \tag{9.3}$$

with,

$$\gamma = 1 + \sum_{k=1}^{m} e^{\left(\sum_{j=1}^{x} - \tau_{ij}\right) + k \cdot (\beta_v - \delta_i)} \text{ and } \mathrm{P}\left(\mathrm{a}_{vi} = 0 \mid \beta_v, \tau_{ij}\right) = \frac{1}{\gamma}$$

The justification of measures was based on empirical evidence of data meeting the requirements of measurement as set out by RMT [43]. This included:

1. Assessment of fit of observed item responses to expected responses by means of tests of fit that provide approximately chi-square-distributed statistics [64, 74]; these fit statistics were applied at an adjusted sample size of 500 to avoid excessive power of the test of fit and deviations from the chi-square-distribution;
2. Assessment of local independence of items by means of residual correlations, which should be close to zero, ensuring that all items contribute equally and independently to the total score [25, 53, 89];

Table 9.2 Characteristics of the complete sample for instrument calibration and the sample of cigarette users for the crosswalk (quantitative study)

	Complete sample (instrument calibration) frequency (%)	Cigarette users (crosswalk sample) freq(%)
Total sample size	2434	250
Age group		
18–34 years	460 (18.9)	29 (11.6)
35–49 years	814 (33.4)	63 (25.2)
50+ years	1160 (47.7)	158 (63.2)
Gender		
Male	1460 (60.0)	103 (41.2)
Female	974 (40.0)	147 (58.8)
Other	0.0 (0.0)	0 (0.0)
Educational attainment		
High school or less	330 (13.6)	67 (26.8)
Some college/college degree	966 (39.7)	110 (44.0)
Bachelor degree or more	1138 (46.8)	73 (29.2)
Race/ethnicity		
White	1972 (81.0)	219 (87.6)
Non–white	462 (19.0)	31 (12.4)
User group		
Exclusive user	1181 (48.5)	250 (100.0)
Poly user	1253 (51.5)	0 (0.0)
	Mean (S.D.)	Mean (S.D.)
Current use of cigarettes (number/day)	5.2 (8.3)	12.5 (8.6)
ABOUT–Dependence raw scores		
Extent of use (0–10, reversed)	5.0 (2.9)	5.5 (2.6)
Signs and symptoms (0–20)	10.5 (5.3)	11.3 (4.5)
Behavioral impact (0–20)	6.3 (5.1)	5.9 (4.4)
ABOUT–Dependence Index (0–50)	21.8 (11.5)	22.7 (9.7)
FTND raw score (0–10)	4.2 (2.5)	3.7 (2.3)

3. Assessment of unidimensionality by means of principal component analysis of item residuals, expecting that correlations of residuals are all random and no common underlying component exists [38, 73];
4. Assessment of item invariance by tests for differential item functioning for user type, TNPs, and various sociodemographic variables, which examine whether the observed responses only depend on the person location and not on third variables [18, 39];

5. Assessment of reliability, defined as the proportion of true variance in the total variance of person measures, which is captured by the person-separation-index in RMT [3]; and
6. Assessment of targeting by means of a graphical matching of item and person locations (targeting plot) including the information curve relevant for uncertainty [15, 40].

Based on the qualitative studies six items were suspected to be redundant, that is their content was mirrored in six other items. For these pairs of items, local independence was not expected to hold. It was further assumed that one item was problematic potentially resulting in misfit. For the final item-reduced scale(s), a minimum level of reliability (person-separation-index) of at least 0.7 was deemed acceptable, while 0.8 or more was considered desirable.

9.4.6 Results of Psychometric Analyses

Since the initial analysis of all 19 items showed some indication of a lack of strict unidimensionality, the conceptualization was reconsidered, and, eventually, three domains of self-reported dependence were proposed:

- *Extent of use*, covering the timing, urgency, or pervasiveness with which the product[s] is/are used;
- *Behavioral impact*, comprising the behavioral aspects of dependence and its impact on daily activities; and
- *Signs and symptoms*, related to the perception of symptoms of dependence experienced by TNP users.

Analyses within each domain confirmed redundancy where it had been expected and also identified an item as inadequate that was suspected to misfit. The item-reduced version consisted of 12 items (see Table 9.3 for item locations and fit statistics, and Table 9.5 for descriptive item statistics) still providing full content coverage from a qualitative view as only redundant items had been eliminated. Reliability was acceptable (0.73) for the two-item extent-of-use scale but higher than 0.80 for the five-item behavioral-impact scale (0.81) and the five-item signs-and-symptoms scale (0.89). Table 9.3 summarizes the psychometric properties of the ABOUT–Dependence self-report instrument. The hierarchy of the items is reflected in Table 9.3 by the order in which items are listed. For signs and symptoms, low dependence is represented by a strong desire, followed by the feeling that completely quitting is difficult, while high dependence is associated with the impression that it is hard to control the need or urge to use the TNP. In the behavioral impact domain, the automaticity (use more than intended) marks the lower end of dependence, while avoiding activities altogether where one could not use the TNP, is at the upper end.

A noteworthy observation was the fact that the three domain scores correlated between 0.5 and 0.8 in the entire sample raising the question whether a higher-order essentially unidimensional variable could be established by combining the three

Table 9.3 Psychometric assessment of the ABOUT–Dependence instrument

Item (abbreviated)[a]	Response scale	Item location (SE) [thresholds]	Item fit χ^2 (df, p)	Highest residual correlation [with item[a]]	Unidimensionality	Reliability PSI (Cronbach's alpha)
Extent-of-use: Covering the timing, urgency or pervasiveness of product use (scale level fit $\chi^2 = 17.09$, df = 16, p = 0.38)						
[1] How long before going to sleep last product	'0–5 min' to 'more than 3 h'	−0.04 (0.03) [−2.73; –0.95; 0.13; 1.04; 2.33]	4.56 (8, 0.80)	NA	NA	0.73 (0.83)
[2] How soon after you woke up first product	'0–5 min' to 'more than 3 h'	0.04 (0.02) [−1.72; –0.85; −0.20; 0.69; 2.26]	12.53 (8, 0.13)	NA	NA	
Behavioral impact: Behavioral aspects of dependence and its impact on daily activities (scale level fit $\chi^2 = 69.39$, df = 45, p = 0.01)						
[3] Use more than you intended	'Never' to 'all the time'	−0.86 (0.03) [−3.54; –1.66; 0.18; 1.60]	14.39 (9, 0.11)	−0.20 [5]	Person measures from most distinct item sets identified by principal component analysis[b] significantly different at error rate of 5%: 6.1% (CI: 5.1%; 7.0%)	0.81 (0.89)
[4] Stop what you were doing to use product(s)	'Never' to 'all the time'	−0.23 (0.03) [−1.78; –1.28; 0.55; 1.62]	15.41 (9, 0.08)	−0.07 [7]		
[5] Use in situation you weren't sup-posed to	'Never' to 'all the time'	0.14 (0.03) [−1.27; –0.49; 0.61; 1.69]	6.81 (9, 0.66)	−0.20 [3]		
[6] Sneak off use prodct	'Never' to 'all the time'	0.31 (0.03) [−0.74; –0.73; 0.81; 1.90]	10.63 (9, 0.30)	−0.04 [7]		
[7] Avoid activity because couldn't use prod	'Never' to 'all the time'	0.63 (0.03) [−0.50; –0.24; 1.00; 2.28]	22.16 (9, 0.01)	−0.04 [6]		
Signs and symptoms: Related to feelings and experience of dependence by TNP users (scale level fit $\chi^2 = 38.20$, df = 45, p = 0.75)						
			11.38 (9, 0.25)	0.00 [10]		0.89 (0.91)

[8] Strong desire to use product(s)	'Never' to 'all the time'	−0.78 (0.04) [−4.23; −2.45; 0.66; 2.92]			Person measures from most distinct item sets identified by principal component analysis[b] significantly different at error rate of 5%: 3.1% (CI: 2.4%; 3.8%)	
[9] Difficult . . . to completely quit product(s)	'Not at all' to 'extremely'	−0.52 (0.03) [−2.50; −0.96; 0.13; 1.23]	3.12 (9, 0.96)	−0.20 [11]		
[10] Feel that you "HAD to have one"?	'Never' to 'all the time'	0.12 (0.03) [−2.63; −1.27; 1.38; 3.02]	10.53 (9, 0.31)	0.00 [8]		
[11] Need product (s) to function "normally"	'Not at all' to 'extremely'	0.33 (0.03) [−1.64; −0.71; 1.10; 2.57]	5.84 (9, 0.76)	−0.20 [9]		
[12] Hard to control the need or urge to use	'Never' to 'all the time'	0.85 (0.03) [−1.36; −0.23; 1.67; 3.31]	7.35 (9, 0.60)	−0.03 [8]		

[a] The instrument with full item wording is available through MAPI Research Trust (https://eprovide.mapi-trust.org/instruments/about-dependence)

[b] The subsets of items identified by the principal component analysis of the item residuals represent items that are maximally internally consistent within each subset but maximally heterogeneous across subsets. As a result, the subsets represent the most extreme deviation from unidimensionality that is possible in the data

Table 9.4 Psychometric assessment of the ABOUT–Dependence instrument subtest structure

Subtest	Subtest location (standard error)	Item fit χ^2 (df, p)	Highest residual correlation [with subtest #]	Reliability PSI (Cronbach's alpha)
[A] Extent of use: e-cigarettes	−0.31 (0.03)	8.89 (9, 0.45)	−0.35 [F]	0.82 (0.83)
[B] Extent of use: cigars, waterpipe	0.10 (0.04)	11.65 (9, 0.23)	−0.42 [E]	
[C] Extent of use: cigarettes, smokeless tobacco, pipes, NRT	−0.05 (0.02)	14.88 (9, 0.09)	−0.39 [F]	
[D] Extent of use: poly users	0.02 (0.01)	4.15 (9, 0.90)	−0.63 [E]	
[E] Behavioral impact	0.35 (0.01)	6.44 (9, 0.69)	−0.42 [B]	
[F] Signs and symptoms	−0.11 (0.04)	20.18 (9, 0.02)	−0.35 [A]	
Scale level		38.22 (54, 0.95)		

domain scores. To this effect, items within each domain were added up to sum scores, which were treated as super-items or subtests. Adequate fit of the data to the model supported a common essentially unidimensional variable of self-reported dependence (Table 9.4). As a higher-order variable, the composite measure, named the ABOUT–Dependence Index, captures the common variance shared by the three domain scores. Any domain-specific variance is attributed to measurement error in this approach. Therefore, it is important to compare the reliability estimate with reliability of the individual scales at the domain-level. Since reliability remained high (PSI = 0.82), the composite measure was justifiable from a statistical point of view. In terms of invariance, the extent-to-use domain required adjustment for differential item functioning by the type of TNP used. This was accomplished by estimating TNP-specific parameters for this subtest. In contrast, the domains of behavioral impact and signs and symptoms demonstrated invariance with respect to the respondents using different TNPs. Thus, a unified metric for self-reported dependence could be defined, allowing for comparisons across user types and different TNPs. In the combined unidimensional measure, signs and symptom items are easier to endorse than items in the behavioral impact domain implying that feelings set in first with behavioral consequences signaling higher dependence, demonstrating that the composite measure also appeared to be conceptually meaningful.

It should be noted that the deviation from strict unidimensionality in the analysis of all items could also be accounted for by applying a multidimensional Rasch model [17, 49, 70]. In the interest of a parsimonious representation of the measurand and the correspondence of the newly developed instrument and the legacy instruments required for establishing crosswalks, we primarily, and successfully, pursued the higher-order variable approach.

The degree to which items match respondents (cigarette users) with regard to the measurand is visualized in the targeting plots for each of the three domains and the ABOUT–Dependence Index (Fig. 9.1). While the distributions of item thresholds (lower part of each plot) and person measurements (upper part) match well for the extent-of-use domain and the signs-and-symptoms domain, there is some mismatch for the behavioral-impact domain suggesting that these items generally indicate higher levels of dependence than those observed among respondents in the sample. This finding is not necessarily a disadvantage as it implies that the behavioral-impact domain captures the higher end of self-report dependence ensuring that the ABOUT–Dependence Index, which combines all three domains, covers the underlying latent variable very broadly. In fact, the targeting of the ABOUT–Dependence Index is very good implying a reliable item threshold estimation, robust fit assessment, and reasonable ranges of uncertainty for the vast majority of respondents.

The validity of measurements by the ABOUT–Dependence instrument has been further substantiated by addressing several aspects of validity. Experts agreeing to the content of the items being highly relevant and matching the conceptual definition of the variable provided evidence of content validity. Internal construct validity has been supported by fit of the data to the Rasch model (see Tables 9.3 and 9.4) and a meaningful hierarchy of items in terms of the amount of dependence the items represent [23, 24]. Correlations of measurements by the ABOUT–Dependence instrument and measurements from product-specific instruments (ranging between 0.83 and 1.00 when correcting for attenuation due to measurement error) demonstrated convergent validity, that is the correspondence of measurements of the same variable from different self-report instruments. These findings also support construct validity and justify the establishment of crosswalks.

The person location measures, obtained by non-linear transformations of the sum score across all items depending on different user types and TNPs because of differential functioning, are expressed in the same metric and are, thus, comparable. However, the metric is rather technical referring to natural logarithms of odds. Therefore, the linear measures on the logit scale were linearly transformed into a more intuitively interpretable and accessible metric of 0–100 (Table 9.5).

9.4.7 Addressing Traceability: Method

The new self-report instrument and existing self-report instruments measuring dependence are based on the same holistic concept of dependence, comprising behavioral aspects (self-reported use of TNPs), impact of the experience of dependence on daily activities, and perceptions of signs and symptoms of dependence experienced by TNP users. The conceptual correspondence was also supported by convergent validity. It is therefore desirable to establish a common metric of dependence and make measurements based on different instruments comparable.

The alignment of multiple instruments measuring the same measurand, or their equating, has a long research tradition with the Rasch model lending itself due to the invariance property [75, 88]. The Rasch model allows for the construction of test

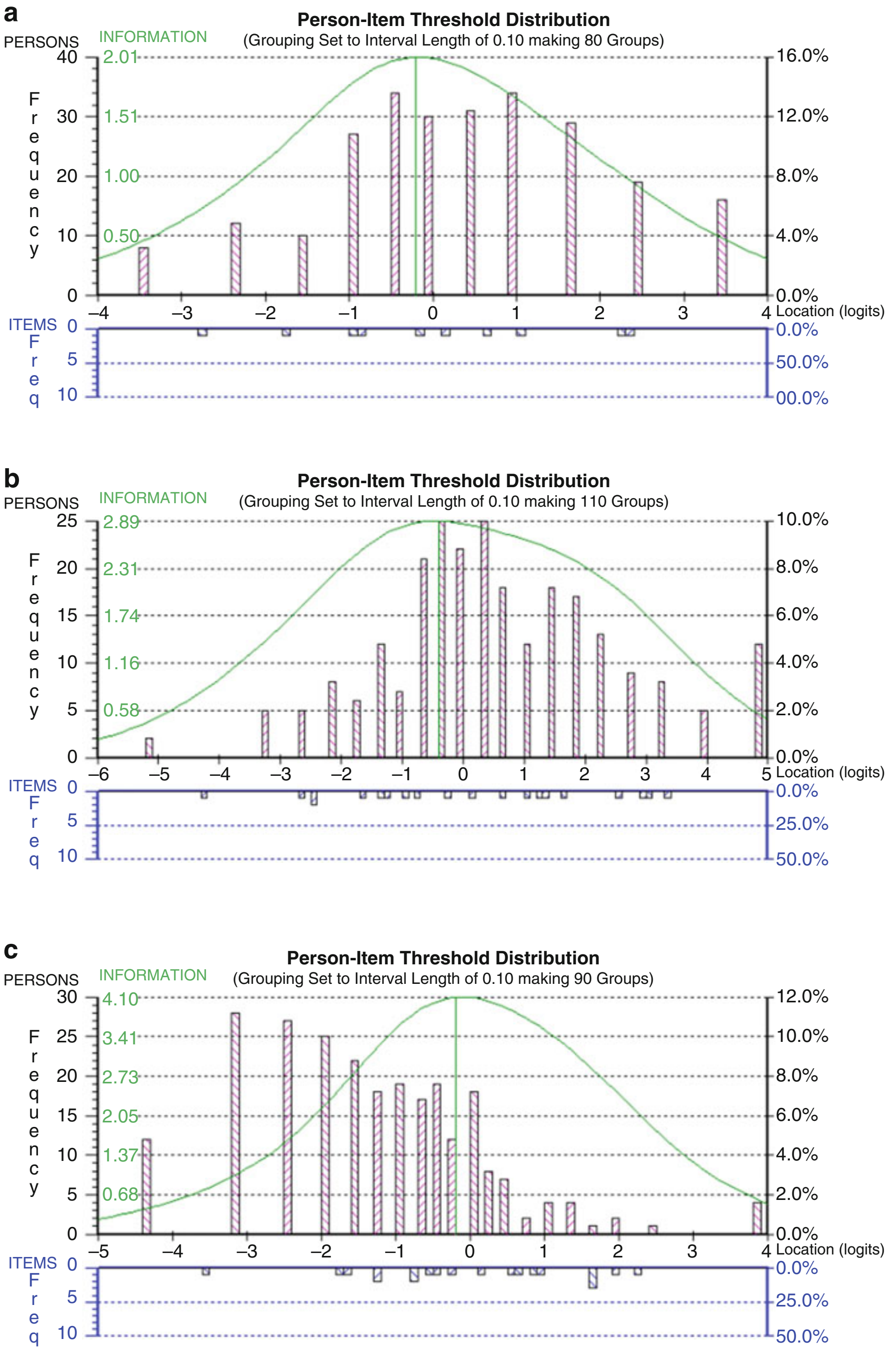


Fig. 9.1 Targeting plots of the three ABOUT–Dependence domains and the ABOUT–Dependence Index for cigarette users. (**a**) Targeting plot of the *extent-of-use* domain. (**b**) Targeting plot of the *signs-and-symptoms* domain. (**c**) Targeting plot of the *behavioral-impact* domain. (**d**) Targeting plot of the *ABOUT-Dependence Index*

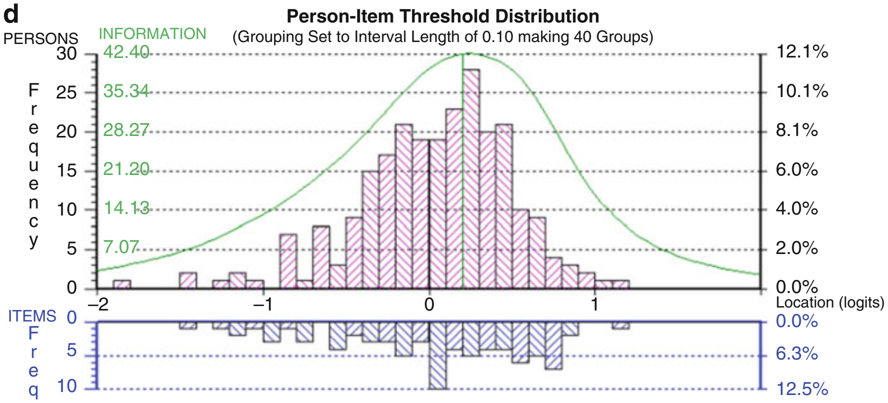


Fig. 9.1 (continued)

Table 9.5 Descriptive statistics of the ABOUT–Dependence items and the ABOUT–Dependence Index measure

Item (abbreviated)#	Complete sample (instrument calibration) n = 2434 Mean (S.D.)	Cigarette users (crosswalk sample) n = 250 Mean (S.D.)
Extent of use		
[1] How long before going to sleep last product (reversed)	3.5 (1.5)	3.7 (1.4)
[2] How soon after you woke up first product (reversed)	3.5 (1.6)	3.9 (1.5)
Behavioral impact		
[3] Use more than you intended	2.8 (1.1)	2.9 (1.0)
[4] Stop what you were doing to use product(s)	2.4 (1.3)	2.4 (1.2)
[5] Use in a situation where you weren't supposed to	2.2 (1.2)	1.9 (1.1)
[6] Sneak off to use product(s)	2.1 (1.2)	1.9 (1.2)
[7] Avoid an activity because you couldn't use product(s)	1.9 (1.2)	1.8 (1.1)
Signs and symptoms		
[8] Strong desire to use product(s)	3.4 (1.0)	3.5 (0.9)
[9] Difficult . . . to completely quit product(s)	3.4 (1.4)	3.8 (1.2)
[10] Feel that you "HAD to have one"?	3.0 (1.2)	3.2 (1.0)
[11] Need product(s) to function "normally"	3.0 (1.3)	3.0 (1.2)
[12] Hard to control the need or urge to use	2.7 (1.2)	2.8 (1.1)
ABOUT–Dependence Index measure (0–100 metric)	52.1 (14.2)	54.0 (10.1)

networks with comparable measurements [30]. Specifically in health measurement, linking instruments by equating is often referred to as a "crosswalk" between instruments. While equating generally means that linear person measures based on parameter estimation in the Rasch model are used, crosswalks typically accommodate the still widespread use of raw scores. A crosswalk converts the raw score on one instrument into an equivalent raw score on another instrument. While this is arguably useful, moving from raw scores to linear measures is a step long overdue. Hence, the illustration of a crosswalk for self-report dependence will not only provide a conversion of different raw scores but, more importantly, establish a common linear metric for different instruments by means of equating.

Two types of equating can be distinguished with either items or persons forming the link in the test design that ensures comparability of the metric [90]. In vertical equating, or common item equating, measures for different persons are placed onto the same metric by administering a common set of items to all persons while adding different items for different groups of persons [48, 63]. Essentially, this means that different instruments partially overlap. Common-item equating is widely used in education. In contrast, horizontal equating, or common-person equating, uses the same persons but mutually exclusive items forming different instruments [56]. Crosswalks are an example of common-person equating. While common-item equating sometimes suffers from targeting problems (the common items are relatively easy for one group of respondents but relatively hard for another), no such difficulties occur in common-person equating provided the instruments to be equated are properly targeted towards the sample. Related to equating but focusing on the unit of measurement, Humphry and Andrich [45] developed a framework for aligning measurements with a different unit. The approach has recently been applied to consumer data in marketing based on differently directed response scales [72].

9.4.8 Addressing Traceability: Establishing a Crosswalk

In the following, the establishment of a crosswalk between the newly developed ABOUT–Dependence instrument and the most widely used legacy instrument, the FTND, is demonstrated. The same can be done for any other existing instrument assessing nicotine dependence.

The crosswalk was based on a co-calibration of responses to the ABOUT–Dependence instrument and to the FTND. The sample was confined to exclusive users of cigarettes, for whom ABOUT–Dependence and FTND measurements were theoretically comparable. For the FTND, a subtest was defined, while the three subtests for the ABOUT–Dependence instrument were retained. In order to preserve the previously established metric, subtest parameter values for the ABOUT–Dependence instruments were anchored to the previously estimated values from the whole study sample. Therefore, only the subtest parameters of the FTND were estimated linking it to the metric established by ABOUT–Dependence. Next, conversions of raw scores to linear measures were retrieved for the ABOUT–Dependence

Table 9.6 Establishing a crosswalk between the ABOUT–Dependence instrument and the FTND for cigarette users (extract)

ABOUT–Dependence raw score	ABOUT–Dependence linear measure (Rasch logit metric)	FTND linear measure (Rasch logit metric)	FTND raw score
0	−2.37		0
1	−1.83		0
2	−1.49		0
3	−1.28	−1.33	0
4	−1.13		0
5	−1.01		1
6	−0.90		1
7	−0.81	−0.84	1
8	−0.72		1
9	−0.64		1
10	−0.57		2
…	…		…

instrument on the one hand and the FTND on the other. Each raw score on the FNTD was then matched with a score on the ABOUT–Dependence instrument in a way that the associated linear measures, which are on the same metric, were as close as possible.

As an illustration, a raw score of 1 on the FTND implied a linear measure on the common metric of −0.84 (see Table 9.6). The closest match on the ABOUT–Dependence instrument was −0.81, which corresponded to a raw score of 7. Therefore, a score of 7 on the ABOUT–Dependence was equivalent to a score of 1 on the FTND. The same was true for all raw scores between 5 and 9 on the ABOUT–Dependence instrument as their associated linear measures were closest to −0.81 on the FTND. Since all ABOUT–Dependence raw scores were also transformed into a linear 0-to-100 metric, FTND scores can also be converted to that metric, which is highly recommended as raw scores are non-linear whereas measures in the 0-to-100 metric are. This is particularly relevant, when establishing further crosswalks to other legacy instruments along the lines of the crosswalk to the FTND. Then the use of the 0-to-100 metric is crucial as is takes differential item, functioning into account. Figure 9.2 shows a graphical representation of the crosswalk. Table 9.7 shows the crosswalk between the FTND and the ABOUT–Dependence instrument.

9.4.9 Comparison of Predicted and Observed Scores on the Two Instruments

The crosswalk translates raw scores on one instrument into raw scores on the other instrument. Consequently, the scores on both instruments can be predicted based on the scores on the other. Table 9.8 shows the FTND raw score predicted from the

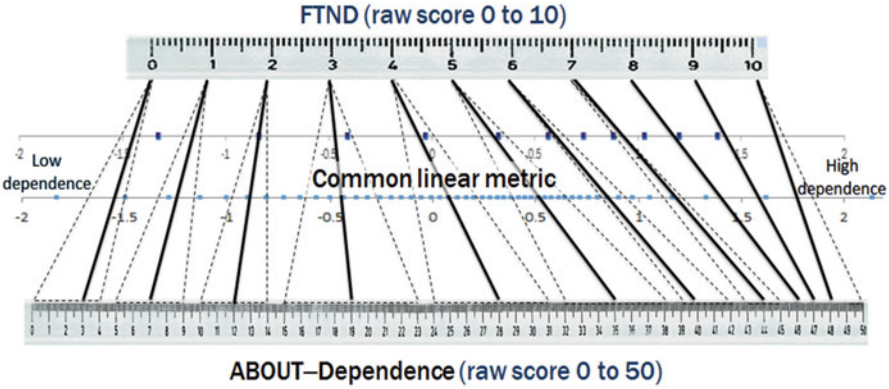


Fig. 9.2 Crosswalk between the ABOUT–Dependence instrument and the FTND for cigarette users

ABOUT–Dependence Index raw score and the crosswalk compared to the actually observed FTND raw scores. The means of the observed score correspond reasonably well with the predicted FTND scores. A regression analysis of the mean observed score on the predicted score reveals a slight regression-to-the-mean effect (unstandardized regression coefficient $\beta_1 = 0.79$), which can be attributed to measurement error in the ABOUT–Dependence Index as measurement error in the FTND is drastically reduced by taking the mean of the observed score. On average, the crosswalk works well, though (Pearson correlation $r = 0.93$ of the predicted score and the mean observed score). At the individual level, the relationship is weaker ($r = 0.69$ of the predicted FTND raw score and the individually observed FTND score), which is due to measurement uncertainty in the FTND. These findings need to be interpreted with caution, though, as the available sample sizes for expected scores of 0 and scores beyond 6 are small.

Conversely, Table 9.9 compares the ABOUT–Dependence Index raw score predicted from the FTND raw score with the actually observed ABOUT–Dependence Index raw score. While the means of the observed raw scores generally increase with the predicted score (except at the upper end), the regression-to-the-mean effect is now much stronger (unstandardized regression coefficient $\beta_1 = 0.52$), which is mainly a consequence of the measurement error in the FTND. The observed regression to the mean compromises the prediction of the ABOUT–Dependence Index based on the FTND raw score, even though the predicted score and the mean observed score correlate at $r = 0.98$. At the individual level, the relationship between the predicted ABOUT–Dependence Index raw score and the individually observed ABOUT–Dependence Index score ($r = 0.72$) matches the relationship observed for predicted and observed FTND scores. The regression-to-the-mean effect remains strong, though (unstandardized regression coefficient $\beta_1 = 0.49$).

Table 9.7 Crosswalk between the FTND and the ABOUT–Dependence

FTND (cigarettes) raw score (0 to 10)	Common metric (0-to-100) of self-reported dependence (uncertainty FTND)	Common metric (0-to-100) of self-reported dependence (uncertainty ABOUT–Dependence)	ABOUT–Dependence raw score (0 to 50)	FTND (cigarettes) raw score (0 to 10)	Common metric (0-to-100) of self-reported dependence (uncertainty FTND)	Common metric (0-to-100) of self-reported dependence (uncertainty ABOUT–Dependence)	ABOUT–Dependence raw score (0 to 50)
(0)	0	0 (26)	0	(4)	58	58 (4)	26
(0)	16	16 (13)	1	(4)	59	59 (4)	27
(0)	22	22 (10)	2	4	59 (11)	59 (4)	28
0	27 (20)	27 (9)	3	(4)	60	60 (4)	29
(0)	30	30 (8)	4	(4)	61	61 (4)	30
(1)	32	32 (7)	5	(4)	61	61 (4)	31
(1)	35	35 (7)	6	(5)	62	62 (4)	32[a]
1	36 (16)	36 (6)	7	(5)	63	63 (4)	33
(1)	38	38 (6)	8	(5)	63	63 (4)	34
(1)	40	40 (6)	9	5	64 (10)	64 (4)	35
(2)	41	41 (6)	10	(5)	65	65 (4)	36
(2)	43	43 (5)	11	(5)	65	65 (4)	37
(2)	44 (14)	44 (5)	12	(5)	66	66 (4)	38
2	45	45 (5)	13	(6)	67	67 (4)	39
(2)	47	47 (5)	14	(6)	68	68 (4)	40
(2)	48	48 (5)	15	6	68 (9)	68 (4)	41
(3)	49	49 (5)	16	(6)	69	69 (5)	42
(3)	50	50 (5)	17	(7)	70	70 (5)	43
(3)	51	51 (5)	18	7[a]	71 (9)	71 (5)	44
3	52 (12)	52 (4)	19	(7)	73	73 (6)	45
(3)	53	53 (4)	20	8	75 (10)	75 (6)	46

(continued)

Table 9.7 (continued)

FTND (cigarettes) raw score (0 to 10)	Common metric (0-to-100) of self-reported dependence (uncertainty FTND)	Common metric (0-to-100) of self-reported dependence (uncertainty ABOUT–Dependence)	ABOUT–Dependence raw score (0 to 50)	FTND (cigarettes) raw score (0 to 10)	Common metric (0-to-100) of self-reported dependence (uncertainty FTND)	Common metric (0-to-100) of self-reported dependence (uncertainty ABOUT–Dependence)	ABOUT–Dependence raw score (0 to 50)
(3)	54	54 (4)	21	(8)	77	77 (7)	47
(3)	55	55 (4)	22	9	80 (11)	80 (8)	48
(3)	55	55 (4)	23	10	86 (13)	86 (11)	49
(4)	56	56 (4)	24	(10)	100	100 (18)	50
(4)	57	57 (4)	25				

[a] Lowest range of uncertainty

Table 9.8 Comparing predicted and observed scores on the FTND

	Observed FTND score			
Predicted FTND raw score based on crosswalk	Mean	Minimum	Maximum	n
0	0.5	0.0	2.0	6
1	0.6	0.0	3.0	17
2	2.1	0.0	5.0	34
3	3.1	0.0	7.0	77
4	4.7	1.0	8.0	75
5	6.0	3.0	9.0	26
6	6.8	4.0	9.0	8
7	7.0	6.0	8.0	3
8	5.5	5.0	6.0	2
9	n.a.	n.a.	n.a.	0
10	8.0	8.0	8.0	2

Table 9.9 Comparing predicted and observed scores on the ABOUT–Dependence Index

	Observed ABOUT–Dependence Index score			
Predicted ABOUT–Dependence Index raw score based on crosswalk	Mean	Minimum	Maximum	n
3	10.5	1	23	30
7	15.9	2	25	18
13	18.3	4	30	32
19	19.9	6	34	33
28	23.0	13	39	40
35	26.6	14	46	42
41	30.6	18	47	21
44	31.6	17	43	20
46	37.1	24	50	12
48	36.5	33	40	2
49				

Every sum score on the FTND can be converted to a corresponding sum score on the ABOUT–Dependence instrument by looking up the score in bold in columns 1 and 5, respectively, and reading the ABOUT–Dependence score in the same row in columns 4 and 8, respectively. Vice versa, every sum score on the ABOUT–Dependence instrument in columns 4 and 8, respectively can be converted to an equivalent score on the FTND in columns 1 and 5, respectively (the corresponding score may appear in parentheses). Each sum score on either instrument can also be converted to a measure in the transformed 0-to-100 metric in columns 2/3 and 6/7, respectively. In these columns, values in parentheses state the range of uncertainty implying a 68% confidence interval for the true measure. For a 95% interval, these values need to be multiplied by 1.96 (or 2). For extreme scores (0 and 100), uncertainty is based on extrapolated estimates of the standard error of measurement.

9.4.10 Measurement Uncertainty of Self-Reported Dependence

Crosswalks enable the expression of measurements based on different instruments measuring the same variable in a common metric. That said, the original uncertainty in the measurements is retained. The higher number of items in the ABOUT–Dependence instrument compared to the FTND implies that measurements are associated with lower uncertainty – that is, smaller standard errors of measurement (SEM; [5]). One SEM plus or minus around the estimate of the measure constitutes a 68% interval as the range of uncertainty. The difference in uncertainty has also been exemplified by the fact that, with the exception of a score of 48, multiple raw score values on the ABOUT–Dependence instrument map onto the same score on the FTND. Table 9.7 lists the range of uncertainty for measurements based on the FTND versus the ABOUT–Dependence for each measure on the 0-to-100 metric (SEM stated in parentheses implying a 68% confidence interval when adding and subtracting from the measurement value). Figure 9.3 shows the SEM of measures based on the ABOUT–Dependence versus the FTND for cigarettes. Uncertainty is smallest near the center of the scale (the exact location depends on all item threshold locations) but gets larger when approaching the extremes implying a U-shaped distribution. The lowest uncertainty for the ABOUT–Dependence instrument is at about 60 on the 0-to-100 metric. For the FTND, the lowest uncertainty is reached at about 70 on the common metric. Generally speaking, the ABOUT–Dependence instrument provides measurements with approximately half the uncertainty compared with to FTND. What is more, the lower number of items confines the FTND to the range between 27 and 80.

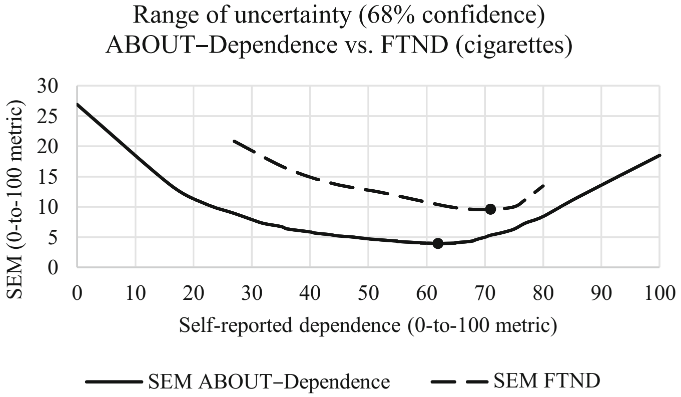


Fig. 9.3 Uncertainty of the ABOUT–Dependence instrument versus the FTND for cigarettes. NOTES: *SEM* standard error of measurement (uncertainty), *FTND* Fagerström Test for Nicotine Dependence. The dot indicates lowest range of uncertainty

9.5 Discussion

The Rasch model for measurement allows for equating different instruments that measure the same measurand. Equating not only provides crosswalks that translate raw scores on one instrument to raw scores on another (a very useful procedure for clinical applications, for example), but even more importantly it establishes a common linear metric of the measurand. The metric becomes essentially independent of the instruments, which are consolidated to form a network of instruments. In other words, a crosswalk as such is a rather pragmatic tool, while the common metric represents a theoretical advance. A metrologically-situated approach to equating therefore goes well beyond ensuring comparability. A common framework for multiple measurement instruments may also facilitate the development of a more powerful conceptual theory of the measurand transcending a particular instrument.

Notwithstanding the theoretical and practical challenges measurement standards in metrology pose to the social sciences, adopting the principle of metrological traceability certainly has the potential to propel social measurement to a higher level. In the long run, revealing measurement mechanisms that theoretically explain why instruments work the way they do will be the pinnacle of social measurement. Given the intricacies involved, in the short run, capitalizing on equating different instruments that are supposed to measure the same measurand will help the social sciences realize part of the benefits metrological traceability has to offer.

However, RMT also provides more meaningful standard errors of measurement that allow for tackling the second fundamental principle of metrology, the range of uncertainty in measurement. Since the range of uncertainty is expressed in the same metric as the measurements, which is then common to all instruments connected through equating, uncertainty can be compared across instruments. This allows for a more informed assessment of measurement quality by each instrument.

Caution is required, though, when it comes to the implementation of a common metric for a measurand in the social sciences. It is crucial to establish the metric based on an instrument that allows for measurement that meets the requirements as set out by the Rasch model. New instrument development is the best avenue towards achieving this goal, even though a re-analysis of an existing scale with a sound qualitative underpinning can be a viable alternative. Apart from the scientific advantages in terms of metrological traceability, crosswalks also have the potential to facilitate the dissemination of new instruments. By linking past research, and continued use of legacy instruments, to current studies using the new instrument, a crosswalk provides a smooth transition and, thereby, is likely to raise acceptability and use of the new scale. Conversely, the integration of legacy instruments into newly established networks including the common metric may also mask limitations of existing instruments. So if there are deficiencies in these instruments, it is crucial to disclose them.

The measurement of dependence on TNPs through self-report instruments illustrates the limitations in social measurement and how they can be overcome using RMT. Until recently, measurement of nicotine dependence has been very fragmented with a number of different instruments providing measurements confined to specific TNPs. The increasing habit of using multiple products concurrently poses new problems for measurement of dependence. The development of a new instrument, the ABOUT–Dependence, helps overcome these scientifically challenging problems by providing comparable measurements of nicotine dependence in case of exclusive use of different TNPs or concurrent use of multiple TNPs. On the other hand, especially designed to accommodate cigarette use, the FTND represents a gold standard in the field of tobacco research when it comes to measuring nicotine dependence. Thus, these two instruments are prime candidates for a crosswalk linking them to one another. Further instruments can then be linked to the common metric contributing to a metrological system of nicotine dependence measurement.

The example illustrates how a crosswalk can be established. A possible limitation lies in the relatively small sample size available for linking the FTND to the common metric. While the common metric itself was established based on the entire sample in the quantitative study ($n = 2434$), the FTND could only be linked based on 250 exclusive users of cigarettes. A more serious limitation was revealed by comparing predicted raw scores based on the crosswalk and actually observed raw scores. For individual measures, the range of uncertainty in the FTND is relatively high implying non-trivial discrepancies between predicted scores on the ABOUT–Dependence and actual ABOUT–Dependence scores. This finding is hardly surprising. While the crosswalk establishes a link between instruments, the range of uncertainty in measurements is retained even though the measures are converted to a more elaborate metric. For obvious reasons, the conversion of ABOUT–Dependence scores to FTND scores is less problematic. But it is also less useful as it entails a loss in information. In practice, crosswalks between instruments of similar range of uncertainty are certainly more expedient.

Social sciences are encouraged to modify their research agenda by including the establishment of a common metric across different instruments by means of equating. This is an important step in taking metrological traceability into account. That said, addressing traceability by linking instruments should be considered a transitory technology permitting comparable measurement in the social sciences. In the long run, revealing measurement mechanisms and transforming conceptual theories into quantitative theories is key.

9.6 Conclusion

Measurement in the social sciences is, with respect to fundamental principles of metrology, lagging behind physical measurement. First, the widely used CTT-based measurement models essentially presume measurement and are mostly concerned with group level parameters of purported measurements and their statistical behavior. Second, measurement in the social sciences is based on conceptual theories that mostly are of qualitative nature with some theories at least allowing for a hypothesized ordering of item. In particular, in educational proficiency testing the order of item difficulties is, as a rule, anticipated. Nevertheless, these theories do not reveal a measurement mechanism or fundamental reference that could be linked quantitatively to the empirical properties of an instrument. Rather, the empirical findings inform the quantitative aspects of the conceptual theory. Consequently, third, the social sciences witness a fragmentation of measurement instruments that are supposed to measure the same concept but do not allow for comparable measurements. Traditional approaches for aligning scores from different instruments are cumbersome and require large representative samples. Thus, the state-of-affairs is highly unsatisfactory from the perspective of metrological traceability.

The status quo of social measurement, therefore, implies two fundamental challenges: first, the measurement theory is to be advanced to provide justifiable and comparable measurement; and second, more elaborate conceptual theories are to be developed. With respect to the measurement theory, recent progress in psychometric modelling has helped narrow the gap between physical and social measurement. Rasch measurement theory provides a framework for invariant measurement, which is a prerequisite for meaningful measurement at the individual level and straightforward comparability. In contrast to these universally applicable psychometric innovations, advances in the formation of more sophisticated conceptual theories apply to specific applications. What is more, today's psychometrics results from decades of innovative research, while advances of conceptual theories required by theory-based traceability are, by comparison, a much more recent phenomenon.

Equating has a long tradition in measurement based on the Rasch model. In the past, though, it has primarily served pragmatic purposes, for example by linking multiple forms of a test to a common metric in education. In health, crosswalks help convert one raw score into another. A metrologically-situated approach to equating means that the method of equating is used to link multiple instruments to one another, while also establishing a common metric that transcends instruments. Measurements based on different instruments become traceable to one another and the range of uncertainty of a specific measurement becomes exposed, too, in a comparable way. This approach is an important step towards metrological traceability in social measurement.

Acknowledgments This work has been part of scale development study to which the following collaborators have contributed (listed in alphabetical order): Linda Abetz-Webb, Esther Afolalu, Stefan Cano, Christelle Chrea, Karl Fagerström, Jed Rose, and Rolf Weitkunat.

References

1. N.D. Adroher, A. Tennant, Supporting construct validity of the evaluation of daily activity questionnaire using linear logistic test models. Qual. Life Res. **28**(6), 1627–1639 (2019)
2. D. Andrich, A rating formulation for ordered response categories. Psychometrika **43**(4), 561–573 (1978)
3. D. Andrich, An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. Educ. Res. Persp. **9**(1), 95–104 (1982)

4. D. Andrich, A general form of Rasch's extended logistic model for partial credit scoring. Appl. Meas. Educ. **1**(4), 363–378 (1988a)
5. D. Andrich, *Rasch Models for Measurement (No. 68)* (Sage, Newbury Park/London, 1988b)
6. D. Andrich, B.S. Sheridan, G. Luo, *Rumm2030: Rasch Unidimensional Measurement Models [Computer Software]* (RUMM Laboratory Perth, Western Australia, 2009-2012)
7. D. Andrich, Advances in social measurement: A Rasch measurement theory, in *Perceived Health and Adaptation in Chronic Disease: Stakes and Future Challenge*, ed. by F. Guillemin, A. Leplège, S. Briançon, E. Spitz, J. Coste, (Routledge, Milton Park/New York, 2018), pp. 66–91
8. C.L. Backinger, P. Fagan, M.E. O'Connell, R. Grana, D. Lawrence, J.A. Bishop, J.T. Gibson, Use of other tobacco products among US adult cigarette smokers: Prevalence, trends and correlates. Addict. Behav. **33**(3), 472–489 (2008)
9. P.J. Batterham, M. Sunderland, T. Slade, A.L. Calear, N. Carragher, Assessing distress in the community: Psychometric properties and crosswalk comparison of eight measures of psychological distress. Psychol. Med. **48**(8), 1316–1324 (2018)
10. S. Bell, Measurement good practice guide no. 11. A beginner's guide to uncertainty of measurement, in *Teddington, Middlesex, United Kingdom: National Physical Laboratory*, (Issue 2, 2001). Retrieved from https://www.dit.ie/media/physics/documents/GPG11.pdf
11. N.L. Benowitz, Nicotine addiction. N. Engl. J. Med. **362**(24), 2295–2303 (2010)
12. BIPM, IEC, IFCC, ILAC, IUPAC, IUPAP, ISO, OIML, *The International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM), 3rd Edition, Version with Minor Corrections, JCGM 200*, (2012)
13. I. BIPM, I. IFCC, I. ISO, IUPAP, O, Evaluation of measurement data – Guide to the expression of uncertainty in measurement. Technical Report No. JCGM 100: 2008 GUM 1995 with minor corrections, in *Joint Committee for Guides in Metrology*, (2008)
14. T.R. Black, *Doing Quantitative Research in the Social Sciences: An Integrated Approach to Research Design, Measurement and Statistics* (Sage, Thousand Oaks, 1999)
15. T. Bond, Z. Yan, M. Heene, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (Routledge, New York, 2020)
16. R.F. Bornstein, S.C. Rossner, E.L. Hill, M.L. Stepanian, Face validity and fakability of objective and projective measures of dependency. J. Pers. Assess. **63**(2), 363–386 (1994)
17. D.C. Briggs, M. Wilson, An introduction to multidimensional measurement using Rasch models. J. Appl. Meas. **4**(1), 87–100 (2003)
18. J. Brodersen, D. Meads, S. Kreiner, H. Thorsen, L. Doward, S. McKenna, Methodological aspects of differential item functioning in the Rasch model. J. Med. Econ. **10**(3), 309–324 (2007)
19. K.A. Bronnikov, V.D. Ivashchuk, M.I. Kalinin, V.N. Mel'nikov, V.V. Khruschov, On the choice of fixed fundamental constants for new definitions of the SI units. Meas. Tech. **59**(8), 803–809 (2016)
20. T.L. Caputi, E. Leas, M. Dredze, J.E. Cohen, J.W. Ayers, They're heating up: Internet search query trends reveal significant public interest in heat-not-burn tobacco products. PLoS One **12**(10), e0185735 (2017)
21. R.S. Caraballo, L.L. Pederson, N. Gupta, New tobacco products: Do smokers like them? Tob. Control. **15**(1), 39–44 (2006)
22. H. Chang, *Inventing Temperature: Measurement and Scientific Progress* (Oxford University Press, New York, 2004)
23. C. Chrea, T. Salzberger, L. Abetz-Webb, E.F. Afolalu, S.J. Cano, J. Rose, R. Weitkunat, K.O. Fagerström, Development of a tobacco and nicotine products dependence instrument, in *Poster Presented at the Society for Research on Nicotine and Tobacco (SRNT) 24th Annual Meeting, Baltimore, USA*, (2018a)
24. C. Chrea, T. Salzberger, L. Abetz-Webb, E.F. Afolalu, S.J. Cano, J. Rose, R. Weitkunat, K.O. Fagerström, Development of a fit-for-purpose tobacco and nicotine products dependence instrument, in *Poster Presented at the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Europe, Barcelona, Spain*, (2018b)

25. K.B. Christensen, G. Makransky, M. Horton, Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. Appl. Psychol. Meas. **41**(3), 178–194 (2017)
26. B.G. Dodd, W.R. Koch, Item and scale information functions for the successive intervals Rasch model. Educ. Psychol. Meas. **54**(4), 873–885 (1994)
27. O.D. Duncan, M. Stenbeck, Panels and cohorts: Design and model in the study of voting turnout, in *Sociological Methodology*, ed. by C. C. Clogg, (American Sociological Association, Washington, DC, 1988), pp. 1–35
28. J.O. Ebbert, C.A. Patten, D.R. Schroeder, The Fagerström test for nicotine dependence-smokeless tobacco (FTND-ST). Addict. Behav. **31**(9), 1716–1721 (2006)
29. S.E. Embretson, S.P. Reise, *Item Response Theory* (Psychology Press, New York/London, 2013)
30. G. Engelhard Jr., D.W. Osberg, Constructing a test network with a Rasch measurement model. Appl. Psychol. Meas. **7**(3), 283–294 (1983)
31. K.O. Fagerström, Measuring degree of physical dependence to tobacco smoking with reference to individualization of treatment. Addict. Behav. **3**(3–4), 235–241 (1978)
32. K. Fagerström, Determinants of tobacco use and renaming the FTND to the Fagerström Test for Cigarette Dependence. Nicotine Tob. Res. **14**(1), 75–78 (2011)
33. G.H. Fischer, *Einführung in die Theorie psychologischer Tests [Introduction to the Theory of Psychological Tests]* (Huber, Bern, 1974)
34. G.H. Fischer, Derivations of the Rasch model, in *Rasch Models, Foundations Recent Developments, and Applications*, ed. by G. H. Fischer, I. W. Molenaar, (Springer, New York, 1995), pp. 15–38
35. W.P. Fisher Jr., A.J. Stenner, Theory-based metrological traceability in education: A reading measurement network. Measurement **92**, 489–496 (2016)
36. J. Foulds, S. Veldheer, J. Yingst, S. Hrabovsky, S.J. Wilson, T.T. Nichols, T. Eissenberg, Development of a questionnaire for assessing dependence on electronic cigarettes among a large sample of ex-smoking E-cigarette users. Nicotine Tob. Res. **17**(2), 186–192 (2015)
37. P. Giacomo, The new definition of the meter. Am. J. Phys. **52**(7), 607–613 (1984)
38. P. Hagell, Testing rating scale unidimensionality using the principal component analysis (PCA)/t-test protocol with the Rasch model: The primacy of theory over statistics. Open J. Stat. **4**(6), 456–465 (2014)
39. C. Hagquist, D. Andrich, Recent advances in analysis of differential item functioning in health research using the Rasch model. Health Qual. Life Outcomes **15**(1), 181–188 (2017)
40. C. Hagquist, M. Bruce, J.P. Gustavsson, Using the Rasch model in nursing research: An introduction and illustrative example. Int. J. Nurs. Stud. **46**(3), 380–393 (2009)
41. C. Häne, T. Sattler, M. Pollefeys, Obstacle detection for self-driving cars using only monocular cameras and wheel odometry, in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (IEEE, 2015), pp. 5101–5108
42. T.F. Heatherton, L.T. Kozlowski, R.C. Frecker, K.O. Fagerstrom, The Fagerström test for nicotine dependence: A revision of the Fagerstrom tolerance questionnaire. Br. J. Addict. **86**(9), 1119–1127 (1991)
43. J. Hobart, S. Cano, Improving the evaluation of therapeutic interventions in multiple sclerosis: The role of new psychometric methods. Health Technol. Assess. (Winchester, England) **13**(12), iii, ix–x, 1–177 (2009)
44. R.R. Holden, Face validity, in *The Corsini Encyclopedia of Psychology*, ed. by I. B. Weiner, W. E. Craighead, (Wiley, Hoboken, 2010)
45. S.M. Humphry, D. Andrich, Understanding the unit in the Rasch model. J. Appl. Meas. **9**(3), 249–264 (2008)
46. P. Jha, Avoidable global cancer deaths and total deaths from smoking. Nat. Rev. Cancer **9**(9), 655–664 (2009)
47. G. Karabatsos, The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. J. Appl. Meas. **2**(4), 389–423 (2001)

48. H. Kelderman, Common item equating using the loglinear Rasch model. J. Educ. Stat. **13**(4), 319–336 (1988)
49. H. Kelderman, Multidimensional Rasch models for partial-credit scoring. Appl. Psychol. Meas. **20**(2), 155–168 (1996)
50. O. Kivinen, T. Piiroinen, The relevance of ontological commitments in social sciences: Realist and pragmatist viewpoints. J. Theory Soc. Behav. **34**(3), 231–248 (2004)
51. S. Kreiner, K.B. Christensen, Person parameter estimation and measurement in Rasch models, in *Rasch Models in Health*, ed. by K. B. Christensen, S. Kreiner, M. Mesbah, (ISTE Limited, London/Hoboken, 2013), pp. 63–78
52. F. M. Lord, M. R. Novick (eds.), *Statistical Theories of Mental Test Scores* (Addison-Wesley, Reading, 1968)
53. I. Marais, Local dependence, in *Rasch Models in Health*, ed. by K. B. Christensen, S. Kreiner, M. Mesbah, (ISTE Limited, London/Hoboken, 2013), pp. 111–130
54. A. Markou, Neurobiology of nicotine dependence. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. **363**(1507), 3159–3168 (2008)
55. G.N. Masters, A Rasch model for partial credit scoring. Psychometrika **47**(2), 149–174 (1982)
56. G.N. Masters, Common-person equating with the Rasch model. Appl. Psychol. Meas. **9**(1), 73–82 (1985)
57. J.A. Maxwell, K. Mittapalli, Realism as a stance for mixed methods research, in *Sage Handbook of Mixed Methods in Social & Behavioral Research*, ed. by A. Tashakkori, C. Teddlie, (Sage, Thousand Oaks, 2010), pp. 145–168
58. J. Melin, L.R. Pendrill, S.J. Cano, E.M.P.I.R. NeuroMET, Towards patient-centred cognition metrics. J. Phys. Conf. Ser. **1379**(1), 012029 (2019) IOP Publishing
59. J. Michell, Quantitative science and the definition of measurement in psychology. Br. J. Psychol. **88**(3), 355–383 (1997)
60. J. Michell, *Measurement in Psychology – A Critical History of a Methodological Concept* (Cambridge University Press, Cambridge, 1999)
61. I.M. Mills, P.J. Mohr, T.J. Quinn, B.N. Taylor, E.R. Williams, Redefinition of the kilogram: A decision whose time has come. Metrologia **42**(2), 71 (2005)
62. E. Muraki, Information functions of the generalized partial credit model. Appl. Psychol. Meas. **17**(4), 351–363 (1993)
63. T.R. O'Neill, J.L. Gregg, M.R. Peabody, Effect of sample size on common item equating using the dichotomous rasch model. Appl. Meas. Educ. **33**(1), 10–23 (2020)
64. J.F. Pallant, R.L. Miller, A. Tennant, Evaluation of the Edinburgh post-natal depression scale using Rasch analysis. BMC Psychiatry **6**(1), 28–37 (2006)
65. M. Pérez-Ríos, M.I. Santiago-Pérez, B. Alonso, A. Malvar, X. Hervada, J. de Leon, Fagerstrom test for nicotine dependence vs heavy smoking index in a general population survey. BMC Public Health **9**(1), 493–497 (2009)
66. D. Rogosa, Casual models do not support scientific conclusions: A comment in support of freedman. J. Educ. Stat. **12**(2), 185–195 (1987)
67. G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests* (Danish Institute for Educational Research, Expanded Edition (1980) with Foreword and Afterword by B.D. Wright. Chicago: The University of Chicago Press, Copenhagen, 1960)
68. G. Rasch, On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements in symposium on scientific objectivity, Vedbaek. Dan. Yearb. Philos. **14**, 58–94 (1977)
69. B. Rodu, W.T. Godshall, Tobacco harm reduction: An alternative cessation strategy for inveterate smokers. Harm Reduct. J. **3**(1), 1–23 (2006)
70. J. Rost, The growing family of Rasch models, in *Essays on Item Response Theory*, Lecture Notes in Statistics, Vol 157, ed. by A. Boomsma, M. A. J. van Duijn, T. A. B. Snijders, (Springer, New York, 2001), pp. 25–42
71. T. Salzberger, *Measurement in Marketing Research: An Alternative Framework* (Edward Elgar, Northampton, 2009)
72. T. Salzberger, M. Koller, The direction of the response scale matters–accounting for the unit of measurement. Eur. J. Mark. **53**(5), 871–891 (2019)

73. E.V. Smith Jr., Understanding Rasch measurement: Detecting and evaluating the impact of multidimenstionality using item fit statistics and principal component analysis of residuals. J. Appl. Meas. **3**(2), 205–231 (2002)
74. R.M. Smith, Fit analysis in latent trait measurement models. J. Appl. Meas. **1**(2), 199–218 (2000)
75. R.M. Smith, G.A. Kramer, A comparison of two methods of test equating in the Rasch model. Educ. Psychol. Meas. **52**(4), 835–846 (1992)
76. A.J. Stenner, M. Smith III, Testing construct theories. Percept. Mot. Skills **55**(2), 415–426 (1982)
77. A.J. Stenner, M. Smith III, D.S. Burdick, Toward a theory of construct definition. J. Educ. Meas. **20**(4), 305–316 (1983)
78. A.J. Stenner, W.P. Fisher Jr., M. Stone, D. Burdick, Causal Rasch models, Front. Psychol. 4 Article 536 (2013) 1–14. https://doi.org/10.3389/fpsyg.2013.00536
79. S.S. Stevens, On the theory of scales of measurement. Science **103**, 667–680 (1946)
80. M. Stock, The watt balance: Determination of the Planck constant and redefinition of the kilogram. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **369**(1953), 3936–3953 (2011)
81. M.H. Stone, Fisher's information function and Rasch measurement. J. Appl. Meas. **9**(2), 125–135 (2008)
82. M.H. Stone, B.D. Wright, A.J. Stenner, Mapping variables. J. Outcome Meas. **3**(4), 308–322 (1999)
83. D.H. Taylor Jr., V. Hasselblad, S.J. Henley, M.J. Thun, F.A. Sloan, Benefits of smoking cessation for longevity. Am. J. Public Health **92**(6), 990–996 (2002)
84. R.E. Traub, *Reliability for the Social Sciences, Theory and Applications*, Sage Measurement Methods for the Social Sciences (Sage, Thousand Oaks, 1994)
85. C.A. Velozo, K.L. Byers, Y.C. Wang, B.R. Joseph, Translating measures across the continuum of care: Using Rasch analysis to create a crosswalk between the functional independence measure and the minimum data set. J. Rehabil. Res. Dev. **44**(3), 467–478 (2007)
86. E. Versi, "Gold standard" is an appropriate term. BMJ: Br. Med. J. **305**(6846), 187 (1992)
87. M. Wilson, Seeking a balance between the statistical and scientific elements in psychometrics. Psychometrika **78**(2), 211–236 (2013)
88. E.W. Wolfe, Equating and item banking with the Rasch model. J. Appl. Meas. **1**(4), 409–434 (2000)
89. W.M. Yen, Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Appl. Psychol. Meas. **8**(2), 125–145 (1984)
90. C.H. Yu, S.E. Osborn-Popp, Test equating by common items and common subjects: Concepts and applications. Pract. Assess. Res. Eval. **10**(4), 1–19 (2005)
91. A.H. Zwinderman, Pairwise parameter estimation in Rasch models. Appl. Psychol. Meas. **19**(4), 369–375 (1995)

Chapter 10
The Role of Construct Specification Equations and Entropy in the Measurement of Memory

Jeanette Melin and Leslie R. Pendrill

Abstract Memory ability, together with many other constructs related to disability and quality of life, is of growing interest in the social sciences, psychology and in health care examinations. This chapter will focus on two elements aiming at understanding, predicting, measuring and quality-assuring constructs with examples from memory measurements: (i) explicit methods for testing theories of the measurement mechanism and establishment of metrological standards and (ii) substantive theories explaining the constructs themselves. Building on entropy as a principal explanatory variable, analogous to its use in thermodynamics and information theory, we demonstrate how more fit-for-purpose and valid memory measurements can be enabled. Firstly, memory task difficulty, extracted from a Rasch psychometric analysis of memory measurements of experimental data such as from the European NeuroMET project, can be explained with a construct specification equation (CSE). Based on that understanding, the CSE can facilitate the establishment of objective and scalable units through the generation of novel certified reference "materials" for metrological traceability and comparability. These formulations of CSEs can also guide how best to compose new memory metrics, through a judicious choice of items from various legacy tests guided by entropy-based equivalence, which opens up opportunities for formulating new, less onerous but more sensitive and representative tests. Finally, we propose and demonstrate how to formulate CSEs for person ability, correlated statistically and clinically with sets of biomarkers, that can be a means of providing diagnostic information to enhance clinical decisions and targeted interventions.

Keywords Entropy · Construct specification equations · Short term memory · Attention span

J. Melin (✉) · L. R. Pendrill
Research Institutes of Sweden, Gothenburg, Sweden
e-mail: jeanette.melin@ri.se

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3_10

10.1 Introduction

Typical responses to tests of memory, and similar constructs, inform decisions of conformity but, as raw scores, the responses are challenging to handle since they lie on ordinal scales, and often lack construct theories and established metrological standards. In response to a call for consensus methods and procedures to enable measurable quantities for constructs such as memory in clinical tests and examinations [9] we focus in this chapter on:

(i) explicit methods for testing theories of the measurement mechanism and for the establishment of metrological standards for the measured constructs deduced from response data from for example memory tests and examinations; and
(ii) substantive theories explaining the constructs themselves.

Differences between these – the quantity as measured compared with the quantity itself, respectively – reflect imperfections in the measurement system used, expressed as measurement errors and measurement uncertainties.

The methods and theories presented in this chapter build on pioneering work by Stenner and his colleagues (e.g., [64–68]). Like those authors, we adopt Rasch measurement theory (RMT) [56] as a key technique of handling raw scores in preparation for formulation of construct specification equations. But, in the interests of advancing quality assurance of the relevant constructs, we examine here the concept of causality further as well as adopt a more explicit measurement system approach. To draw on the full benefits of analogies with engineering measurement system analysis (MSA) [3], we make a clearer distinction between the responder – regarded as a "measurement instrument" – and the test item construct associated with the measurement object than originally proposed by Rasch [56, 57].

After a brief section on the definitions of terms, the chapter starts with a short description of the measurement system approach. Thereafter, the main part of the chapter presents the rational for CSEs as substantive construct theories, including contrasting our approach with related work, and the role of entropy in a metrological framework both for each construct – such as person ability and task difficulty – of interest in itself as well as constructs explicitly connected with the measurement mechanism, i.e., quantities as measured. The chapter closes with an account where those explicit methods are applied, together with substantive theories in the context of memory measurements, to overcome some of the shortcomings of current neuropsychological assessments and to ensure that quantities are traceable as far as possible to metrological standards and are metrologically legitimated. Illustrations are provided with two tests of short-term memory: a non-verbal tap recalling test (*Corsi Block Test* (CBT), [13]) and a verbal digit recalling test (*Digit Span Test* (DST), [74]). The chapter is then closed by addressing some limitations and providing a summary and outlook for the future where we highlight the possibility of formulating CSEs based on entropy for memory task difficulty to design new, less onerous but more sensitive and representative tests largely by "cherry-picking the best" items from existing batteries of cognitive tests, such as exemplified with CBT and DST.

10.2 Definitions

There are many different terms used in the literature such as observed outcomes, rating scores, counts etc. when referring to test responses. To avoid any conflation and/or misunderstanding we will use the term raw scores throughout this chapter to denote test responses which are classifications (assignments $P_{success}$ to categories, e.g., pass or fail or 1–5 ratings); that is, the observed outcome before measurement restitution. Raw scores are thus distinct from the measurement outcome which is the measurement result after both observation and restitution [53].

The term "examination" is used here in the sense common in medical care and is considered a synonym for "inspection" as defined for example in ISO/IEC 17000: 2004 §4.3, as including a *determination of conformity (of the entity being inspected) with specified requirements*. In a typical medical examination, a clinician in making a judgment will often weigh together several factors, including for instance the results of memory tests but also anamnesis, biomarkers etc. The term "testing" is the *determination of one or more characteristics of an object (entity) of conformity assessment*, according to the same standard (§4.2). "Examination" and "testing" are thus not synonyms for "measurement" [16] since a "measurement" (on any scale) does not necessarily involve a questioning or an assessment of conformity.

10.3 Methods for Testing Theories of What Is Being Measured

The present section has two aims. The first aim is to posit explicit methods for testing theories of the measurement mechanism in for example memory tests and examinations and for the establishment of measurement standards (etalons) in a metrological framework in such tests by adopting measurement system analysis (MSA) (ASTM 2012). MSA is the classic approach in engineering to describe indirect measurements, in the sense that an operator usually needs some kind of instrument to get a measure of the attribute of the object of interest. In both psychometric and psychophysical measurement systems, a human being – an "agent" – can be regarded as acting as a measurement instrument [49, 50]. In this sense, the MSA approach accords well with the (dichotomous) RMT:

$$\theta - \delta = log\left(\frac{P_{success}}{1 - P_{success}}\right) \tag{10.1}$$

and statements about specific objectivity by Rasch ([58], p. 5) himself:

> The parameters δ_j signify manifestations of a certain property of a set of "objects" which are investigated by means of a set of "agents" characterized by the parameters θ_i. Thus in principle the δ_j stand for properties of the objects per se, irrespective of which θ_i might be

> used for locating them. Therefore they really ought to be appraised without any reference to the θ_i actually employed for this purpose – just like reading a temperature of an object should give essentially the same result whichever adequate thermometer were used.

The MSA approach focusses on quantities as measured which in general differ from the same quantities in themselves – e.g., for the object of interest – owing to imperfections in the measurement mechanism which leads to measurement errors and uncertainties. Further description of MSA, including the process of restitution of the measurand (i.e., the measurement outcome) from the response (i.e., raw scores), can be found in the accompanying chapter in this book by Pendrill and Melin [53] which focusses on person-centred care but draws explicitly on the benefits of analogies between RMT and approaches which are more widely applicable and established in the engineering measurement community.

The second aim of this section is to posit theories which substantively explain constructs, which will take up the remaining and major part of this section.

We start our discussion by examining the concept of causality, being careful as far as possible to present separate descriptions of the measurement mechanism and of the constructs. Thereafter, presentations will be given of how *Construct Specification Equations* (CSEs) support substantive construct theories and enable metrological links to *Certified Reference Materials* (CRMs). This section concludes with a detailed description of the formulation of CSE.

10.3.1 Different Levels of Causality in the Measurement Mechanism and in Constructs

In their seminal paper, Stenner et al. [67] introduce what they term a *"causal Rasch model"*, which in their words: "may be seen as formalizing how a measurement mechanism and an attribute measure cooperate to produce (cause) the observed outcome." Here we examine this causality further in terms of three distinct relations:

The first relation holds that raw scores, for instance from memory tests, can be transformed mathematically onto interval scales. In the words of Linacre and Wright [36]:

> The mathematical unit of Rasch measurement, the log-odds unit or "logit", is defined prior to the experiment. One logit is the distance along the line of the variable that increases the odds of observing the event specified in the measurement model by a factor of 2.718..., the value of "e", the base of "natural" or Napierian logarithms used for the calculation of "log-" odds. All logits are the same length with respect to this change in the odds of observing the indicative event.

Salzberger [60] wrote later:

> The Rasch model tests whether an a priori absolutely scaled raw score represents an a posteriori ... non-linear raw score, which can be transformed into a linear interval-scaled measure of the latent variable.

In our opinion, this first transformation is *not* an expression of causality, but rather a **purely mathematical relation** (in line with [36]) between raw scores and test attributes. That mathematical relation provides the formula when restituting quantitative estimates of the measurand from the ordinality of the raw score responses [52].

A second relation, between raw scores and differences between test attribute variables – such as task difficulty δ_j and person ability θ_i – which RMT enables to be estimated separately and conjointly (Eq. 10.1), is a kind of causal relation but a special kind, which **describes the measurement mechanism**. In line with MSA, causality in this second relation has to do with how measurement information propagates from the measurement object (e.g., a memory task), through the instrument (the person tackling the task), to the system observer, who in a memory test can be a clinician [49]. However, because of the ordinality in the response to the tests considered here – such as memory tests – we refrain from calling the raw scores the dependent variables in a cause-and effect relation since the observed outcomes are in themselves strictly not quantities (i.e., measurable properties).

A third relation – which we consider as expressing a stricter kind of causality than the first two relations above – is between **an RMT-derived attribute as dependent variable and one or more explanatory** (independent) variables, **X** (Eq. 10.2 below). Examples of this are constructs such as memory task difficulty δ_j and person memory ability θ_i explained in terms, respectively, of test item characteristics and person characteristics and biomarkers. Such relations – termed *Construct Specification Equations* (CSE) by Stenner and Smith [64] – are not in the first hand to do with the measurement mechanism but are used to describe the attributes associated with the object or entity in themselves. The next section describes this in more detail.

To sum up, with a CSE as a kind of substantive construct theory addressed in the third kind of strictly causal relation above, the object or entity in itself does not need to be associated with a measurement mechanism. However, as examples of special cases of CSEs, we can also formulate theories of the attributes associated with measurement, that is, of the attributes associated with various elements of a measurement system. In addition to Rasch's [56] emphasis on the importance of providing separate estimates of the attribute of the object being measured (e.g., memory task difficulty) and the instrument being used to make the measurements (e.g., person memory ability corresponding to an attribute of the instrument used for measurement), one should be careful to distinguish construct theories which address the object itself from those describing an element of the measurement system being used in the process. Previous measurement models – such as developed since the 1980s by Stenner et al. – chose to call a test item for an "instrument" and a human responder as an "object". Rasch's original [56] description emphasized that the choice of "object" and "instrument" was agnostic since both attributes appear symmetrically (apart from a change of sign) in the Rasch formula. This flexibility is also present in our own MSA-based approach [49, 50] but to get the full benefit of analogies with engineering MSA it is preferable to describe the human responder as the instrument rather than the item, as cited in the previous section.

Two additional comments: Firstly, notwithstanding that CSEs represent a stricter causality than RMT, construct and measurement theories are handled here as far as possible with equal priority: we do not share the view of others, paraphrasing

Michell ([47], pp. 75–77), that the *instrumental task of quantification, i.e., construction of measurement devices and instruments, is* secondary *to the scientific task of discovering quantitative structure*. We give equal attention here to variations in item scores and in person scores, which is not new. As early as the 1920s, Thurstone stated: "*there is at present a wide but artificial break between the group of men who work in psychophysics with the traditional stimuli and those who attempt to measure educational and social values with little interest in psychophysical theory. . . . It is our hope again to unify the efforts to measure social values with the advancement in psychophysical theory"* [70].

10.3.2 Construct Specification Equations (CSEs) as Substantive Construct Theories

In the words of Stenner et al. [67]: "*Rasch analysis, absent construct theory and an associated specification equation, is a black box in which understanding may be more illusory than not.*" Therefore, when formulating an overall attribute construct to be determined in a quality assured manner, various quality characteristics of the entity of interest need to be identified, described, measurable, predictable and prioritized. In the present context of memory measurement, a construct theory must attempt to explain both memory task difficulty and person memory ability in terms, respectively, of a variety of explanatory variables, such as trial length, entropy and biomarkers.

10.3.2.1 Construct Specification Equations (CSEs) and Validity

Richard Feynman wrote on his last classroom blackboard: *"What I cannot create, I do not understand"*. Substantive construct theories – with which constructs can be "created" – have been described in the literature as keys to demonstrating the crucial notion of validity which refers to the conformity of observations with the intended goal of measurement. For this conformity to be achieved, the construct must exist, and variations in it must cause reproducible variations in the observations taken as evidence supporting measurement [1, 6]. In contrast to those works, again here we stress the importance of distinguishing between the construct itself and the construct as measured [52].

For testing the existence and variation of a construct, CSEs are considered to be at the highest level of construct validation available for social, psychological and health measurements, and correspond to Feynman's ambition to create and understand. The CSEs provide a more specific and rigorously mathematical and causal conceptualization of item attributes (e.g., task difficulties) and person characteristics (e.g., person abilities) than any other construct theory. The relationships between measured quantities in the memory tests illustrated later in this chapter, e.g., from easy to difficult tasks or from low to high abilities of persons, can be explained in terms of the "something" – i.e., the explanatory variables – that causes each variation.

Early work on CSE stem from Fischer's [18] and Prien's [54] studies of mathematics abilities, Latimer's [34] application of Fischer's linear logistic test model (LLTM) to reading ability, and Wright and Stone's [77] and Stenner and Smith's work in the 1970s and 1980s on the Knox Cube Test (KCT) and the Peabody Picture Vocabulary Test [64, 65]. The CSE concept has been advanced considerably in education science in the context of reading comprehension [66], whilst it has largely remained an undiscovered country in health care and other areas. Commercial measures linking reading ability with text complexity in assessments and instruction are the domain of the most well-known example of CSE [66] which uses a 2-variable equation to explain text difficulty: the log mean frequency, based on the sum of frequencies of all words in the same word family in use in written or oral communications, and the log of the text's mean sentence length as a proxy for syntactic complexity [63]. In the next section, *Theories explaining what is being measured*, you find more discussion of construct theories, particularly those based on the concept of entropy.

Extending our introduction to causality, CSE relations have precedence in terms of causality over the two other relations (link functions in GLM and the measurement mechanism in IRT) when one seeks to express how the constructs of interest are to be understood, predicted, measured and quality assured, that is, "created" in Feynman's words. Only when one has the strict causality that a CSE describes [52], can it be said that one has a relation in a form similar to the relations between (so-called derived) quantities in Physics. An example of the latter derives the quantity "force" from two quantities, namely "mass" and "acceleration", where the quantity relation (also a CSE) in this case follows Newton's physical law of motion as an expression of causality. But such universal relations are rare: most CSEs, while objective and causal to a certain extent, will usually not enjoy the universality of relations such as Newton's laws (where the latter apply to all bodies throughout the universe and on all scales, from the microscopic to cosmological), but are instead expressions of limited local validity, analogous to engineering sensor relations and the weaker objectivity characteristic of the social sciences [52]. Together with this, as emphasized in the introduction, again, it is important to distinguish between quantities as such and quantities as measured. Statistics can be applied to either of these, respectively, describe errors/anomalies in quantities (e.g., local inconsistences in object attributes) as well as measurement errors and uncertainties arising from imperfect measurement systems. That is, not all of statistics is measurement-related. For a discussion on descriptive purposes of statistical modelling and the prescriptive purposes of measurement modelling see the accompanying chapter by Fisher.

In a critique, Kyngdon [33] claimed that: *'the Rasch model is not conjoint measurement because the Rasch model 'simply map[s] a set of real numbers (probabilities) into another set of real numbers (differences between logarithmic unit parameters).'* While we agree – as stated above – that the Rasch model can be regarded as a purely mathematical relation, RMT is nevertheless also a uniquely *metrological* approach where attributes for item/object and agent/instrument *are* linked conjointly. At the same time, we do not consider relations in Physics – such as the associative relation between density, volume and mass, or the fundamental law of Nature: Force = Mass x Acceleration – to be “conjoint measurement”, as claimed by Kyngdon [33] and Luce and Tukey [37]. Such relations – such as Newton’s 2nd law or the formula for a strain gauge – admittedly might form the basis of a sensor in some cases but do not in our opinion uniquely define conjoint measurement *per se*. Such relations are in the first case amongst quantities in themselves, irrespective of whether they have been measured or not.

Indeed, when formulating CSEs, the choices of explanatory variables, **X**, (Eq. 10.2 below) are not restricted to observed quantities. Rather, various combinations of variables should be explored as guided by our understanding of the construct and by multivariate analyses. Two examples given in this chapter are, respectively: (i) formulation of explanatory variables based on the concept of entropy as a measure of information [40] and (ii) identification of principal components of variation, based on a principal component analysis (PCA), which are various combinations of attributes and which better explain the construct than these directly observed quantities themselves.

In the literature, the CSE approach in psychometrics has mostly been used and advocated in item response tests in relation to explaining items’ levels of difficulty. However, only when synthetic constructs created from theory enable the consistent and reliable prediction of both memory item (difficulty) and person memory (ability) location calibrations, can we claim to fully understand what our memory test measures. Thus, the role of CSE should not be considered only as a matter for task difficulty, as in the pioneering work [18, 34, 54, 64, 65, 77]. The CSE approach indeed adds further to metrology – as promised in the Introduction – and can be applied in principle to all elements of the measurement system (i.e., the object, the instrument, the operator, the environment and the method). The examples provided in our third section, *Memory measurements* below, will exemplify CSEs for memory task difficulty and person memory ability in experimental case studies.

10.3.2.2 Construct Specification Equations and Construct Modelling

In the light of construct theories, the work by Wilson [75] on *Construct Modelling* is also worth relating to the methods for CSEs in this chapter. Wilson outlined ‘four building blocks’ for measurements and test development: (a) construct map, (b) items design, (c) outcome space, and (d) measurement model. Wilson presents the construct related to the person attribute, e.g., person memory ability, but not the other RMT-derived attribute, viz. memory item difficulty. However, as the measurement outcome for person ability is usually the attribute one wants to use e.g., for health examinations or performance tests, it is natural to start by defining the attribute related to the person and thereafter designing items to be met for persons going up or down the scale of the person attribute. In turn, one should have an ordinal theory with ordering of groups of items and person attributes, respectively [41].

Wilson, while close, did not reach the level of CSE construct validation in his ‘four building blocks’ but did develop this concept together with de Boeck elsewhere [14]. While a CSE mathematically explains the relationship between items, a confirmatory theory tests the ordinal theory by means of empirically estimated item locations [41], which is what Wilson [75] stresses is facilitated via a Wright map's correspondence to the construct map. With a well-developed construct modelling approach, we will gain a better understanding of our measurements, and will also inform the further development of CSE. To advance construct theories and to elaborate CSE, we need a deeper and quantitative understanding associated to each memory task than what is qualitatively considered to be an easy or difficult task. As will be demonstrated in the following sections, task difficulty varies with the degree of order in a task, with less or more entropy, and similarly for different degrees of person ability.

10.3.3 Construct Specification Equations and Metrology

For the quality assurance of any object (a product, service etc.), it is necessary that any measurements of that object are also quality assured [53]. If we are seeking objects with comparable properties, then it is recommended to ensure that the measurements of their properties are also comparable in themselves. Measurement comparability is ensured in metrology, i.e., quality-assured measurement, by traceability to metrological references through calibration.

As said above, CSEs represent the highest level of construct theory. From the metrological point of view, CSEs for task difficulty (an attribute of the object of interest) can be considered to constitute metrological references [52] analogous to '*recipes for certified reference materials (CRMs)*' which provide traceability in fields such as chemistry and materials science. Serving as measurement objects with known values of their characteristic quantities, chemical CRM enable instrument calibration to provide reliable and traceable measurements and are essential as a form of measurement standard in both verification and validation in chemical metrology.

Analogously, in memory measurements, recipes for metrological references (i.e., Rasch estimates for memory task difficulty, δ_j) established through measurand restitution could be formulated in terms of causality as a CSE for memory task difficulty, providing objective and scalable metrological units for traceability. In the same way that chemical CRMs often allow for matrix effects where the surroundings of a chemical component can affect its concentration, the proposed psychometric CRMs could include account of the effects of context. To be qualified as a "certified" reference material (or procedure, RMP), CSEs in any application (such as person-centred healthcare) would need to be subject to requirements analogous to those stipulated for CRM and RMP in analytical chemistry and materials science [25, 27].

In some cases, a calibrated measurement instrument can act as a metrological reference as an alternative to a calibrated measurement object. In the context of memory measurement, this would involve formulation of a CSE for the measurement instrument, in terms of Rasch estimates for person memory ability, θ_j. Elsewhere, we have argued for starting with CSE for task difficulty as CRM is done for practical reasons, analogous to the procedure in quantitative metrology where a metrological standard (etalon) is associated in the first case with an attribute of an object – such as the mass of a weight – rather than the instrument used to measure it, since the latter is arguably less suitable as a metrological reference owing to its complexity and lack of robustness compared with a simple object weight [42].

We are so far refraining from talking about certified reference materials for person ability until we have a valid CSE for task difficulty. Specifically, a CSE for person memory ability can be used for well-designed measurement systems and in turn improve the reliability of task difficult estimates, and a basis for discussions with clinicians about the understanding of the person's memory ability.

10.3.4 Formulation of Construct Specification Equations

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

$$\mathbf{Y} = \sum\nolimits_{k} \beta_k \cdot x_k \tag{10.2}$$

For memory measurements, Rasch estimates, δ_j or θ_j, for each item, *j*, or person, *i*, (Eq. 10.1) 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, $\mathbf{X_k}$.

The explicit identification of the dependent variable follows from cause-and-effect considerations and differs from an implicit function which includes all variables of interest on the same side of the equation. An example of the latter is the Disease State Index for the evaluation of Alzheimer's disease of Mattila et al. [38].

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) [52]: (the programs and algorithms used in this PCR work are detailed in an appendix at the end of this chapter.)

(i) *Principal component analysis (PCA) amongst the set of explanatory variables*, $\mathbf{X_k}$: The initial set, **X**, of explanatory variables in Eq. 10.2 may exhibit correlation, making it unsuitable for direct regression. Principal component analysis (PCA), where a matrix **P** of the principal components 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 principal components of variation are the eigenvectors, **p**, of the covariance of **X**, with eigenvalues λ:

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

(ii) *Linear regression of the Rasch estimates, δ_j or θ_j, against* **X'** *in terms of the principal components*, **P**: As a second step, the Rasch construct **Y** [Eq. 10.1], e.g., task difficulty, δ, or person ability, θ, with ε variation) is expressed

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

by performing a least-squares regression against the principal components:

$$\mathbf{C} = \left(\mathbf{T}^{\mathbf{T}} \cdot \mathbf{T}\right)^{-\mathbf{1}} \cdot \mathbf{T}^{\mathbf{T}} \cdot \mathbf{Y} \tag{10.3}$$

(iii) *Conversion back from principal components (PC) 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:

$$\mathbf{Y_0} = \mathbf{X_0} \cdot \mathbf{P} \cdot \mathbf{C}$$

to yield a linear combination of the explanatory variables, **X**, as shown in Eq. 10.2, where the coefficients in the linear predictor (construct specification equation):

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

Thus, the formulation of CSEs includes two essential multivariate steps, equally applicable and important to explaining memory task difficulty or person memory ability: 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 principal component analysis, form PCR.

The resulting CSE from this PCR analysis provides an important source of information: how much each explanatory variable contributes to explain and predict the variation in the attribute of interest. For instance, in the memory tests exemplified later in this chapter, it will be shown how much variables such as *Entropy, Reversals* and *Average distance* contribute to explaining memory task difficulty. By regressing measured values, **Y**, of the attribute of interest against corresponding predicted values, *zR*, yielded by the CSE, the R^2 index indicates the amount of variance in task difficulty or person ability accounted for by the construct theory. As explained above, the CSE, when combined with RMT, sets forth a measurement theory of raw score variation as well as simultaneously providing the vehicle for confirmation or falsification of the construct theory. Although on an ordinal scale, raw scores as outcomes are still preferred to logits by some users. The Rasch formula (Eq. 10.1) provides a simple means of converting between them.

As will be presented in next section and further illustrated in the third section, our understanding of a construct may suggest combining several quantities to form one major explanatory variable. Entropy, as a measure of order in either task or person and related to the amount of "useful" information, is one actual example explored here. An easier task or a more able person both have more order and less entropy. Once such causality has been recognized, it will be possible not only to formulate CSEs for these constructs but also to argue for the equivalence of items in different tests, with a view to eventually combining these when formulating a new and improved test, with better coverage and less items, as well as linking tests performed in different cultures.

In addition to the above-mentioned PCA when formulating CSEs, it is common to examine the unidimensionality of item attributes also using a PCA, particularly in view of the basic assumption of unidimensionality in RMT. A second kind of PCA is used to examine item residuals of the logistic regression in RMT [35]. Despite the differences between these two kinds of PCA, the results from the two can be expected to yield results which are connected. If there is more than one explanatory variable revealed in the first step of formulating a CSE with PCR, it is also likely to be accompanied by indications of a second dimension in item residual PCA (where the first PC is the primary Rasch attribute). Therefore, one could surmise that, for instance, memory item difficulty might depend on more than one factor – in addition to the degree of order (captured by the entropy term), perhaps a second "dimension" connected with another cause (such as the number of reversals in a sequence), which might scale differently. This is further illustrated below.

All in all, CSE formulation has advantages (familiar from simpler calculations such as forming a mean value and increasing reliability by adding more items or test persons) of reducing measurement uncertainties by including increased numbers of degrees of freedom through combining several explanatory variables as opposed to analyzing each variable singly. The multivariate coefficients determined experimentally (below) have smaller uncertainties than the corresponding relation for each explanatory variable taken univariately.

10.3.4.1 Measurement Uncertainties of Construct Specification Equations

There are always parameters associated with both the object (task) and person attribute values, which characterize the dispersion of the values as indications of doubt in the measured values, i.e., measurement uncertainties. In addition to the measurement uncertainties for each attribute of interest (i.e., $u(\delta)$ for each memory task's difficulty and $u(\theta)$ for each person's memory ability), there two further expressions of measurement uncertainties will be given in this chapter: for each β-coefficient of the CSE (Eq. 10.2) and for the predicted values zR.

10.3.4.2 Measurement Uncertainty for *β*-Coefficients in CSE

It is of interest to express corresponding means and standard uncertainties in the regression β-coefficients of the CSE relating the attribute value, **Y,** (in the examples here: memory task difficulty or person memory ability) to a set of explanatory variables, **X**, as well as statistics for significance testing of various differences amongst attribute values.

The measurement uncertainties in the attribute values (i.e., *u(δ)* for each memory task's difficulty and *u*(θ*)* for each person's memory ability [53] will propagate through the principal component regression described in section 1.4. An initial set of uncertainties in the estimates, $\mathbf{C} = \left(\mathbf{T}^{\mathbf{T}} \cdot \mathbf{T}\right)^{-\mathbf{1}} \cdot \mathbf{T}^{\mathbf{T}} \cdot \mathbf{Y}$ (where $\mathbf{X}' = \mathbf{T} = \mathbf{X} \cdot \mathbf{P}$ and **P** is the matrix of principal components) of the coefficients from the present least-squared analyses, e.g., for the CBT exemplified below, is

$$\widehat{\mathbf{C}} = \begin{bmatrix} 1,95 & 0,74 \\ -0,75 & 1.23 \\ -0,1 & 0.3 \end{bmatrix}$$

where the second column indicates the (expanded, $k = 2$) uncertainties, $\mathbf{U}(\mathbf{C}) = \boldsymbol{k} \cdot \boldsymbol{u}(\mathbf{C})$, in each coefficient (first column) for the three principal components of variation for memory task difficulty, *δ*, and three explanatory variables, *Entropy, Reversals* and *Average distance*, respectively.

To account for the possible effects of heteroscedasticity in the uncertainties, *u(δ)* or *u*(θ*)*, in the attribute values which are different for the different tasks or persons [11, 30], one approach is to use a standard, weighted least-squares fit in which Eq. 3 is replaced by:

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

where the weighting matrix $W_{i,j} = \frac{1}{u_i^2}; i = j; 0, \text{otherwise}$. In the cases studies presented later in this chapter, heteroscedasticity in the different memory tasks and person abilities uncertainties is mostly small, so possible bias and scatter from this is expected not to be a dominant effect.

10.3.4.3 Measurement Uncertainty in the zR Yielded from the CSE

The above-mentioned uncertainties, $\mathbf{u}\left(\widehat{\mathbf{C}}\right)$, in the least-squared coefficient estimates will propagate to produce uncertainties in the CSE (i.e., the linear predictor, Eq. 10.2), where the CSE coefficient expression $\beta = \mathbf{P} \cdot \mathbf{C}$ used when transforming back from PCs **P** to the original explanatory variables **X**.

Corresponding uncertainties in the zR yielded from the CSE will be given by the combined (standard, $k = 1$) uncertainties from each β-coefficient in Eq. 10.2:

$$u(\mathbf{Y}) = \sqrt{\sum_k u(\beta_k)^2 \cdot (x_k)^2}$$

10.3.5 Construct Specification Equation: Ad-hoc or Pre-defined?

In an ideal world, ordinal theories should be adopted in an early phase of measurement construction. However, as stated by Fisher and Wilson [20]: *'if data are available but were generated via a process omitting construct definition and item design, application of a measurement model could still be a useful initial step in beginning to define a construct'*. In the case of memory measurements such as the classic Knox Cube Test (KCT) [32] where block tapping sequences are to be recalled (similar to the CBT exemplified in this chapter), Knox initially specified different lengths of block tapping sequences to match the anticipated achievements of the mental age of each child, i.e., he assumed that longer sequences were more difficult and persons able to manage longer sequences had a higher degree of mental capacity. However, this is neither the case of all memory tests nor a complete means of developing a CSE, but it can serve as a basis for further developing ad-hoc explanatory variables to be used in a CSE. For instance, Knox considered the children's age to be a key in terms of their performance which can guide identification of further person factors related to how mental and memory capacity develop. Similarly, any known pathology may explain memory decline and should be included in a substantive CSE theory.

An example of how a pre-defined understanding of how explanatory variables for memory item difficulty can be used is conditional likelihood estimation of component parameters specified with a linear logistic test model (LLTM) of the relationship between item difficulty and component weights (i.e., explanatory variables). For a memory measurement, in LLTM, memory item difficulties are a linear function of the number and difficulty of the cognitive operations. For instance, Green and Smith [21] formulate LLTM models of KCT memory task difficulties as: $\delta_j^* = -\left(\sum_k \beta_{j,k} \cdot x_k + \varepsilon_\delta\right)$, where x_k denotes the difficulty of cognitive operation (e.g., tapping sequence) k, and $\beta_{j,\,k}$ is the number of times each cognitive operation occurs in item j. A fit of the data to the model can be examined by comparing Rasch model estimates of item difficulties, δ_j, with those estimated by calculating difficulties from the component difficulty estimates, δ_j^*, using a maximum likelihood statistic. The factor ε_δ is regarded as a centering constant, i.e., $\varepsilon_\delta + \frac{1}{L} \cdot \left(\sum_j \sum_k \beta_{j,k} \cdot x_k\right) = 0$. Green and Smith [21] report little difference for the KCT in the estimated component difficulties from their LLTM and the regression results of Stenner and Smith [64], where in both models X_1 is the number of taps; X_2 the distance; and X_3 the number of reversals.

Another example of how pre-defined qualitative explanatory variables can be used for supporting construct validity, although beyond memory measurements, is Adroher and Tennant's [1] recent work in activities of daily living. Seven properties of the items in Evaluation of Daily Activity Questionnaire (EDAQ) were identified and rated in ordinal scales by 39 Occupational Therapists worldwide. Aggregated metric estimates – the weights used to predict item difficulties in LLTM – were derived from the ratings using seven cumulative link mixed models. Parallel to this, classic Rasch estimations of task difficulty, δ_j for each item, j, were assessed and compared with the predictions from the LLTM. In turn, a combination of a theoretical and empirical model enhances the understanding of what is being measured and what causes reproducible variations.

10.4 Theories Explaining Attributes of Interest

In this second section we focus on information theory and entropy as an explanatory variable, **X**, in CSEs. As explained above, when formulating CSEs (Eq. 10.2), experimentally observed quantities characteristic of the items or person may be combined to form new quantities which capture our understanding of the constructs – such as task difficulty and person ability in memory tests. If our understanding is correct, these combinations will turn out to explain variations better than just using the observed quantities, as can be verified experimentally. Here we consider in particular the concept of entropy in describing the causality linking factors determining the effects on each construct.

We start this section by providing a generic conceptualization of entropy and its link to measurement uncertainties and the propagation of information in a measurement system. This is then followed by illustrations of how entropy theoretically and causally relates to constructs such as memory task difficulty, where a more ordered task (less entropy) is expected to be easier, and person memory ability, where a more "ordered" person (less entropy) is expected to be more able. Experimental tests of these theories are then subsequently exemplified for memory examinations.

10.4.1 Entropy in General Terms

Entropy is simply and generally described as a measure of order related to the amount of ”useful” information or "useful" energy [10]: higher entropy implies less order giving an increased risk of more uncertainties due to loss of information or energy, and vice versa, lower entropy implies higher order and less uncertainties.

Entropy was conceived, though not named or expressed mathematically, in thermodynamics by Lazare Carnot [10] who was concerned 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'* ([10], p. 255, §279). All mechanical processes in a closed system involve some irreversible energy loss. Therefore, entropy (i.e., the loss of useful energy) will always increase over time according to what is known now as the second law of thermodynamics, as formulated later by Kelvin [31] and Clausius [12].

Since World War II, entropy has also been a central concept in the field of information theory [62]. For instance, the performance of a communication system depends on how well information (analogously to Carnot's energy) is "usefully" transmitted – lost or distorted – which can be described in terms of informational entropy. In this context, like energy loss in thermodynamics, entropy in communication systems tends to increase over time and useful information is lost. A measurement system is a kind of communication system where each element bears a certain amount of information.

Apart from the rattling machines of the early Industrial Revolution or complex communication systems, we consider how well any task of a certain difficulty is performed by a person with a certain ability in terms of entropy. Entropy enables not only descriptions but also explanations and predictions of performance. As illustrated further below, how well a memory test is performed is considered by taking the measurement mechanism as information and communication described in terms of entropy. "Useful" measurement information – paraphrasing Carnot [10] – can be lost (and sometimes gained) at each of the three main stages – object, instrument and operator – identified in a measurement system as a communication system. Thus, entropy can increase overall through distortion and loss of information anywhere in the measurement system.

Rasch's models for measurement set the stage for our focus here on entropy and information. A broader perspective on entropy in the formulation of concepts is given by Fisher [19].

10.4.2 Entropy, Measurement Uncertainties and Validity

As described above, higher entropy is associated with more loss of "useful" energy or information. In turn, this implies less information and greater uncertainties. Even though measurement uncertainty is more closely related to reliability, as often coupled together with validity, more loss of information and greater measurement uncertainties also give rise to questions about the validity of the measurement. Ultimately the aim of any measurement theory is to compensate as far as possible for the effects of imperfections in the measurement process, in order to obtain the most faithful measure of the quantities of interest. As such, substantive theories explaining the constructs being measured are key to claiming their validity. As said in the Introduction, measurement theories and construct theories are equally important but should also be considered separate issues.

Entropy can be deployed quite generally when extending the probabilistic theory for treating measurement error and uncertainty. In fact, entropy has the advantage of being applicable to all kind of scales, including nominal and ordinal scales, and not only to probability theory but also other inferences in terms of plausibility and belief. Error and uncertainty estimates in qualitative measurements can be expressed in terms of the distortion, fuzziness and lack of clarity of a wide range of characteristics using basic measures of information content, while undistorted, unambiguous and clear patterns and figures are characterized by low values of entropy. Hence, entropy is a useful concept for quality-assurance throughout the measurement process, from stimulation to restitution. Information content can range from basic examples, such as the number of elementary symbols, to increasingly sophisticated information, through syntax, semantic and pragmatic aspects of meaningful information in many contexts.

10.4.3 Entropy, Measurement System and Rasch Measurement Theory

The passage of information through the measurement system can be described in terms of the well-known conditional entropy expression for the corresponding communication system [52]:

$$H(Y|Z) = H(Z, Y) - H(Z)$$

the entropy in the response (Y) when observing the quantity (Z) as the joint entropy reduced by the entropy associated with the measurement object (A) prior to measurement.

The link between a probabilistic description [59] and a corresponding entropy-based approach is based on the informational ‘Shannon’ entropy, H, of any ‘message’ of probability P being proportional to $log(P)$ [62]. This formulation captures the fact that the less expected a message is (i.e., smaller P), the greater the amount of information conveyed (‘surprisal’). Taking the logarithm also facilitates addition and subtraction of different amounts of information.

The “Shannon” entropy terms can thus be expressed in terms of the probability distributions associated with each variable in the measurement process, where these probabilities are multiplied in the expression $P(z, y, z_R) = P(z) \cdot P(y| z) \cdot P(z_R| z, y)$ using the notation used by Rossi in his probabilistic model of the measurement process [59]); and R denotes restitution.

The above expression states how the amount of information changes during transmission in a measurement system. At the start of the measurement process, there is an initial ‘deficit’ in entropy (i.e., ‘surplus’ information) coming from prior knowledge, $H(Z;A)$, of the measurand (attribute, Z, of entity A), again using a notation analogous to that used by Rossi in a probabilistic model of the measurement process [59]. Losses and distortions $H(Y,Z)$ from imperfections in the measurement process increase the entropy, leading finally to a posterior distribution (Q) with entropy $H(Y|Z)$ as the result of the measurement process.

This implies further, in the measurand restitution of memory measurements (i.e., with the Rasch probabilistic formula, Eq. 10.1), that the odds ratio of succeeding with a task, is simply equal to sums and differences of entropy and this connection can be done with the expression:

$$log\,(P) = log\,(\lambda) = log\,(k) - log\,(h)$$

where the average probability P is taken equal to the Poisson distribution factor λ, and h and k enter the equation for the conditional entropy expression above. As recalled by Pendrill and Melin [53], an early form of the psychometric Rasch model (RMT) posits that the odds ratio of successfully performing a task is equal to the ratio of an ability, h, to a difficulty, k [57].

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

which can be written as Eq. 10.1 where the test person ability, $\theta = log\,(h)$, and task difficulty, $\delta = log\,(k)$, can be evaluated by logistic regression to the score data in terms of the probabilities $q\ (P_{success})$. Rasch [60] was keen to point out that his "discovery was a somewhat intuitive achievement . . ., with no relation to any actual item analysis problem" or any analogy to the laws of Physics.

The assumptions behind RMT need of course to be tested, such as the separability of two independent variables contributing to an overall effect or response, which have been questioned over the years, e.g., by Luce and Tukey [37] in their work on simultaneous conjoint measurement. This is done in our case using a battery of tests, based on analyses of variance and regression fit residuals, as exemplified below.

Apart from task difficulty and instrument ability, there is an additional factor – called in IRT the *discrimination* of the person responding – as can be captured in some IRT models with a second and/or third parameter, viz the finite resolution, ρ, of the instrument (i.e., the person taking the test). This discrimination term [24] can be modelled as a change in task difficulty (or corresponding expression for person ability): $\Delta\delta = -(\rho - 1) \cdot \delta$, in turn giving an additional entropy term $H(Z,Y) \sim \Delta\delta = \ln(\sqrt{3} \cdot 2 \cdot u)$ which relates discrimination to measurement uncertainty, u, (here for the case of a uniform distribution) in the response (Y) when observing the quantity (Z). As also discussed in Pendrill and Melin [53], the steepness of the characteristic ogive curve of the item response (in the binary case where one classification – zero, say – goes over to one) is determined by the uncertainty of the person making the classification for a given task. That steepness in turn determines the width – that is, the measurement uncertainty – in each experimental estimate of construct attributes such as task difficulty and person ability, where each estimate is not sharp but broadened to a width $u(\delta)$ and $u(\theta)$, respectively. (As in traditional metrology, provided measurement uncertainties and heteroscedasticity are small, this will not in general hinder the establishment of metrological references with RMT.)

With links to "useful information" and to causality in mind, a task will be easier if there is some degree of order and a poorly performing person can be explained in terms a lack of coordination explained in terms of increases in entropy. Conversely, a task will be more challenging when there is less order and a well performing person can be explained in terms of good coordination. This is further discussed in the next sections for memory task difficulty and person memory ability, respectively.

10.4.4 Entropy to Explain Memory Tasks

The forward sequences of *Corsi Block Test* (CBT) [13], and *Digit Span Test* (DST) [74] will be used as examples. In CBT the person is asked to reproduce the same sequences with different tasks with increasing length and the DST requires a similar recall task, but instead of a tapping sequence the participants are asked to recall digit sequences. As stressed above, the 'something' that causes variation in memory task difficulty needs to explain why some memory items are easier to perform than others.

To include entropy in CSE formulation in memory measurements, the first step is to attempt to explain memory tasks in terms of entropy. The different memory tasks to be recalled – for both non-verbal taps and verbal digits – can be characterized in general in terms of a message in which a number, N_j, ($j = 1,\ldots,J$) of symbols of J different types (taps or digits) can be distributed in a number, G, of categories (or cells) $G = \sum_{j=1}^{J} N_j$. The probability of encountering the jth symbol is $p_j = \frac{N_j}{G}$, which can be summed to unity. According to Léon Brillouin [7], 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 $P = \frac{G!}{\prod_{j=1}^{J} N_j!}$. In turn, the information theoretical entropy, which is a measure of the amount of information in these messages, is then:

$$I = M \cdot lnP = M \cdot \left[\ln(G!) - \sum_{j=1}^{J} \ln(N_j!)\right]$$
$$\cong M \cdot \left[G \cdot \ln(G) - \sum_{j=1}^{J} N_j \cdot \ln(N_j)\right] \qquad (10.4)$$

where M is an arbitrary constant. Stirling's approximation in the final terms of the righthand side of Eq. 10.4 applies when G and N are large, but with modern computer power the approximation is no longer necessary when evaluating the factorial terms. Consequently, this basic expression will be applicable when explaining task difficulty in CBT and DST.

The expression proposed by Brillouin [7] enables entropy to be evaluated for sequences which not only have increasing length (i.e., taps or digits), G, but also some repeats, N, of the same block tapped or digit. In particular, the second entropy term in Eq. 10.4 above shows the expected increase in entropy – and thus the decreased task difficulty – from N repeats.

Other effects which can make a task easier (or more aesthetically pleasing) include some aspect of simplicity, symmetry and the like where the observer recognizes a familiar pattern (similar to seeing the same figure in a digit span test) instead of just a random cloud of dots. In those cases, the extra term in the Brillouin [7] expression shows how the entropy is reduced (making the task easier) by the recognition of a number of symmetric groupings. The difficulty of remembering different block sequences (such as the Corsi block test) was described, for instance by Schnore and Partington [61], in terms of a sum of a set of basic patterns (or chunks) with different information content expressed in terms of entropy and the symmetry of each pattern.

In other work reported separately, we have demonstrated how entropy can explain reductions in task difficulty due to serial position effects (e.g., in word list tests such as R-AVLT) such as primacy and recency [43]. In that work, we propose that a CSE for task difficulty in word learning list tests such RAVLT (length $L = 15$) can be based theoretically on a sum of entropy terms for primary, recency, mid-list recall and word frequency (f) as given, respectively, by:

$$\delta_j = \delta_{j,primacy} + \delta_{j,midrange} + \delta_{j,recency} + \delta_{j,freq} \tag{10.5}$$

where

$$\delta_{j,primacy} = -M \cdot \ln(G_j!); G = item order$$

$$\delta_{j,midrange} = +2 \cdot [M \cdot \ln(\frac{L}{2}!)]$$

$$\delta_{j,recency} = -M \cdot \ln(G_j!); G = L - 1 - item order$$

$$\delta_{j,freq} = -M \cdot \ln f_j$$

and the normalization factor $M = \frac{1}{ln\,(L)}$ [7]. In the case of an odd list length, L, the expression $ln\left(\frac{L}{2}!\right)$ is evaluated by rounding $\frac{L}{2}$ to the nearest integer.

Using the same basic model as in Eq. 10.4, we have recently been able to explain SPE in the immediate recall (IR) of the 15-word learning memory test RAVLT [43].

10.4.5 Entropy to Explain Person Abilities

Having explained task difficulty, we now turn to person ability. The first stage in defining, and testing, explanatory variables for any ability is to formulate an understanding of what causes a low respectively high ability, e.g., is there any known pathology behind? Is it an ability characterized by certain training or learning processes? In the case of measuring person memory ability in Alzheimer's Diseases (AD) spectrum there are some AD-related biomarkers of interest related to amyloid pathology, tau pathology and neurodegeneration [29] as well as brain volumes and other structural parameters related to memory function associated with regions such as the hippocampus.

In general, explanations of persons' abilities are characterized by complexity. This is especially true for cognitively related activities, such as memory abilities as there is an information processing system with 10–100 billion neurons and ~10^{14} synapses in the brain [73]. This system exhibits the highest degree of complexity among all organs in the human body. Connectivity, in particular, is a key to explaining cognitive ability and parallels can be drawn to other high demanding systems in thermodynamic or information theory, and therefore, *Functional Magnetic Resonance Imaging* (fMRI) are gaining more attention in order to explain brain complexity including several entropy-based measures [46, 72].

It should also be noted that a person's memory ability to pay attention to the task to be performed and his or her executive functions may also partly be used to explain the person's ability for tasks requiring recall of a variety of ordered information. There are also potentially other factors within the other elements of the measurement system (i.e., the object, the environment and the method) that can affect the person's memory ability such as how the test is administered (i.e., the method) or if there are surrounding noises and disturbances which might cause stress (i.e., the environment). Such factors should however be separate from the explanatory variables for the person's memory ability in the same way as they are separate from explaining task difficulty, as can be modelled in MSA.

10.5 Examples and Illustrations in Memory Measurements

In this penultimate section, we combine the explicit methods for testing the theories explaining memory item difficulty and person memory ability. We start with a brief review of memory measurements, followed by CSEs based on entropy. It will be demonstrated on the basis of measures of memory task difficulty for two short-term memory recalling tests: CBT and DST, and a CSE for measures of person memory ability will be exemplified based on known biomarkers related to Alzheimer's Disease (AD) spectrum. At the end we summarize the limitations and strengths in the present examples and implications.

10.5.1 Memory Measurements

Since the Ancient Greeks, attempts have been made to understand how the human brain works, such as cognition and mental processes. Today, in the field of 'the measured mind', there are many different person attributes of interest ranging from capacities and abilities to attitudes and personality factors [5]. Our interests are not only in the measured mind but in mind attributes in themselves, irrespective of whether they have been measured. Those interests go well beyond examinations in clinical setting and quantification of health status, such as the field of educational sciences and development psychology.

10.5.1.1 The Broad Picture of Cognition and Mental Processes

Pioneers such as Binet Simon and David Wechsler appeared in the early 20th-century when the first intelligence tests aiming to measure one's underlying mental ability were formulated, including the *Binet Simon Intelligence Scale* and *Wechsler Bellevue Intelligence Scale* [4]. Those tests, as well as similar later intelligence tests, were built on tasks for several aspects of cognition such as recalling, comparing and defining numbers, words and pictures.

Already in the pioneering work, the idea of construct maps (cf. [75]) was present based on the child's age and corresponding longer tests for older children. Likewise, the Knox Cube Test (KCT, similar to the CBT introduced above), introduced at Ellis Island over 100 years ago for the testing of mental limitations in immigrants, had different sequence lengths to be recalled and defined what should be accomplished at different ages. These were structured tests with an objective scoring; either the person could or could not perform the requested task. However, test results were, and still mostly are, based on counts of raw scores, i.e., not measurement outcomes that separate, as in RMT, the raw data into quantitative measures of task difficulty and person ability. In turn, there are often quite large variations in the meaning of score differences and a less than comprehensive understanding of what is being measured.

The idea of constructing tests including several aspects of cognition corresponds to what later became known as a higher ordered construct and Andrich's [2] metaphor of *a rope made up of strands*. This way of scaling tests together requires that the tests included work in a uniform way from less to more to build one construct to be measured on the same scale. In fact, the CSE approach – which reflects how well each construct is understood – can provide an indication of the equivalence of items in different tests, and thus guidance about which equivalent constructs can be reasonably combined to form, hopefully, better and more reliable tests [44]. On the other hand, in the clinical examination, it is common that tests for different cognitive abilities and biological aspects are used parallel, and in turn require that the clinician combine the information to decide diagnosis, drugs and treatment. Similar, this multi-source of information can obviously provide deeper understanding of how cognitive processes occur and relate to each other, while not necessarily explaining a particular construct in itself and what causes its variation.

Moreover, historically, psychological effects in measurement were initially introduced into measurement science [69] in the field of psychophysics during the nineteenth Century, in attempts to relate abstract human sensations to quantifiable physical external stimuli (such as touch pressure, sound pitch). In contrast, psychometrics developed thereafter to include other mental attributes (such as attitude, knowledge, empathy) which are not simply responses to physical stimuli. Tesio [69], in comparing and contrasting the two disciplines, claimed that "psychophysics is deterministic: a cause-effect relationship is assumed between stimulus and response. . .(whereas) the psychometric approach is probabilistic, in that it implies inferences."

Compare Tesio's claim with our description of causality and the responses to entropy. A key enabling insight, in our view, is to connect the treatment of decision risks associated with measurement uncertainty to generalized linear modelling, indeed not only in psychometrics but also across the disciplines. Handling certain more qualitative measurements in the social sciences, psychology and health care examinations in this way unite information theory (the perceptive identification and choice paradigms of psychophysics [28], with a particular focus on the RMT psychometric approach. The idea in psychophysics of modelling responses through the five human senses has been extended to a metrological model of RMT in psychometrics where the human acts as a measurement instrument [49]. But note that, although RMT is also a logarithmic expression (Eq. 10.1), the general linearized model expression in psychometrics is more general than the Weber–Fechner Law of psychophysics, which has a different logarithmic dependence derived in the particular case where a change in the psychometric function is proportional to the fractional change in the stimulus level.

10.5.1.2 Neurodegenerative Diseases and Memory Measures

There are contemporary initiatives calling for biological definitions of neurodegenerative diseases such as Alzheimer's Disease (AD). However, prevailing AD spectrum core criterion are based not only on biomarkers but also on examinations of example cognitive function and ability to function in everyday life [39]. In the clinic, the patient is examined through a combination of history-taking from the patient and a knowledgeable informant together with neuropsychological testing.

Since measures of memory (and other cognitive aspects) currently lack established international standards, it is of course challenging to correctly make fully diagnoses as well as monitor pharmaceutical intervention effects and actual disease progression. In fact, efforts on drugs or therapies delaying or stopping disease progression, in particular in early phases, have mostly been unsuccessful. Such failures, as stressed by Raket ([55], p. 2) *may be due to wrong therapeutic targets or non-efficacious therapies, but it is conceivable that a proportion of trial failure could be attributed to other factors such as study design, endpoints and non-optimal patient populations selection.* To pick up on Raket's first point, about wrong therapeutic targets or non-efficacious therapies: we would emphasize that, without proper measures, it will be challenging to identify the right target needed to be treated. This is especially true in early phases of disease where symptoms (such as memory decline) and signs (such as Amyloid and Tau pathology) are small as well as can be a pre-stage of many different diseases and may not lead to AD spectrum.

To also pick up on Raket's comment on endpoints. There are numbers of widespread neuropsychological tests for different cognitive aspects such as learning and episodic memory; speed and attention; visuospatial functions; language; and executive functions. However, the most commonly used *legacy* neuropsychological tests (e.g., *Mini Mental State Examination* [23]) and *Alzheimer's Disease Assessment Scale-Cognitive Behavior section* [22]) can neither claim accuracy to distinguish between patients (especially in early stage disease due to person-to-item targeting is commonly skewed and there are large measurement uncertainties associated with persons with early memory decline) nor are metrologically legitimated (i.e. lacking metrological references to ensure comparable measurement results) [22, 23, 42]. Despite well-known issues with those tests, the tests are frequently used incorrectly, for instance in studies of biomarker correlations [51].

Of the cognitive aspects related to AD, memory decline is one of the early symptoms. Current AD therapies focus mainly on early-stage disease, which necessitates fit-for-purpose measures to capture early memory decline. Thus, measures of memory decline (or improvement) on individual level needs comparisons at least to specific time points, e.g., annual clinical examinations or longitudinal studies. However, an impaired memory ability can also be established by comparisons e.g., with references values for the same age group or in cross-sectional studies. Both of these comparisons require that the shortcomings of current neuropsychological tests are solved to ensure that the memory quantities are traceable as far as possible to metrological standards and are metrologically legitimated. Thus, this section will illustrate possible solutions in the context of memory measurements based on the explicit methods introduced in the first section and the substantive theories introduced in the second section.

10.5.2 Subjects and Data Analyses

The subjects and data used stem from the project NeuroMET [17] comprising a cohort of 88 subjects with dementia due to suspected AD (n = 26), mild cognitive impairment (MCI) (n = 23) and healthy controls (HC) (n = 39). The mean age was 72 years (range 55–84 years) and 47% were women and 53% men. Of the 88 subjects, 77 had a complete set of memory assessments and biomarkers used for developing CSEs for person memory ability.

During the neuropsychological testing, a correct recall was scored 1 and an incorrect recall was scored 0 for both CBT and DST. This is, however, raw data that needed to be transformed into separate and linear measures for memory task difficulty (δ-parameter) and person memory ability (θ-parameter). To enable this, the Rasch Dichotomous Model was applied to the raw data in the software WINSTEPS®, and consequently used for the formulation of CSEs as described for the two attributes of interest, **Y**, i.e., memory task difficulty and person memory ability.

Motivation of explanatory variables, **X**, for memory task difficulty and person ability is provided, respectively, elsewhere in this Chapter. Explanatory variables for memory task difficulty have been derived based on the variation in the recalling sequences provided; *Entropy* is based on Eq. 10.4, *Reversal* corresponds to the number of times one changes from clockwise to counter-clockwise and the other way around in CBT or counting forward to backward and the other way around in DST, *Average distance* is the average centimeters between blocks in CBT and the average distance between digits in DST [44]. Explanatory variables for person memory ability are well-known AD-related biomarkers, i.e., physical and chemical quantities, for each person in the cohort. Cortical thickness and left hippocampus volume (normalized to each test person's intercranial volume) were obtained by magnetic resonance imaging (MRI) with a 7T scanner and blood-based biomarkers, neurofilament light (NfL), amyloid peptides 1-42/1-40 ratio (Aβ42/40) and total tau proteins (Tau) were measured in plasma [45]. Subsequently, the formulation of CSEs for each attribute and estimations for measurement uncertainties. PTC MathCad Prime 3.1 and its specific modules for the steps in the PCR were used for the analyses and are described in Appendix.

10.5.3 Explaining Memory Task Difficulty

Early CSE work was done in the 1980s by Stenner and Smith [64] and Stenner et al [65] on the Knox Cube Test (KCT) [32]. As explanatory variables, **X**, these early authors were using *Length, Reversals* and *Distance* and, like the present studies, they used Rasch transformation. However, by considering the significance of entropy in measurements, the Brillouin [7] expression for entropy seems in our studies theoretically to be more appropriate. With entropy, it is expected that more ordered sequences will be easier to remember, i.e., the informational entropy (by analogy with thermodynamic entropy) is lower (i.e., more information) when the order of the test (e.g., sequence of blocks or digits) is greater. Thus, the proposal of entropy is in line with studies which showed that inconsistency in performance at longer sequences should be a function of path complexity (as defined by [8]); that memory span is greater for structured than unstructured paths [26, 61]; and that the success rates for different lengths of sequences are overlapping [48], which are properties shared by CBT and DST.

Moreover, in contrast to previous workers who considered the distance between taps as an explanation of KCT task difficulty, the average distance is used below instead. This is to not confound with another explanatory variable – namely the number of taps which enters into the distance – which is already dealt with above. Average distance can be conceptualized as a ratio of signal to noise. Thus, a two-tap sequence such as “2-3” is easier to remember than a two-tap sequence which is spread out over more blocks, such as “1-4”. In the latter case, there is more background noise which the individual must process and filter out.

Below the CBT and DST cases are first handled separately and then related to each other and as a combined measure.

10.5.3.1 Case 1: Tapping Recall

For the CBT, the memory task difficulty values, δ, ranged from -6.5 ± 3.8 to 8.5 ± 3.7 logits where the shortest tapping sequences were easier than the longer. It was, however, evident that the sequence 3-5-1-7-2 was much easier than the other sequence with the same length. The item reliability was 0.95 and fit statistics were satisfactory. In the test situation, there is a possibility for an 8-block-tapping-sequence, but this was eliminated as only one subject achieved that level. Table 10.1 shows the memory task difficulty values, δ, and corresponding explanatory variables used for developing the CSE for CBT.

The above-mentioned PCA when formulating CSEs [section 1.4] indicates firstly that particularly the pair of explanatory variable *Entropy* and *Reversals* are rather

strongly correlated, with the correlation matrix based on the covariance matrix of the explanatory variables indicating a Pearson correlation $\mathrm{r} = \sqrt{R^2} = 0.91$. This is to be expected, since longer series (with consequent lower entropy) will also allow for more reversals to be made in each sequence.

As shown in the following equation, the PCA-based CSE for CBT is dominated by entropy and there is a negligible contribution to the CSEs of *Reversals* and *Average distances*:

$$zR_j = -7(3) + 2.0(7) \times Entropy_j - 0.8(1.2) \times Reversals_j - 0.1(3) \times AveDistance_j$$

Table 10.1 The Rasch estimates, δ_j, for each item, j, and its explanatory variables in CBT

Tapping sequence	Observed memory task difficulty, δ	U (δ)	Entropy	Reversals[a]	Average distance
2 taps, 1st	−6.5	3.8	0.69	0	10.5
2 taps, 2nd	−5.1	2.3	0.69	0	4.7
3 taps, 1st	−3.5	1.5	1.79	0	7.7
3 taps, 2nd	−3.5	1.5	1.79	0	9.9
4 taps, 1st	−1.2	0.8	3.18	0	10.3
4 taps, 2nd	1.1	0.6	3.18	1	9.9
5 taps, 1st	3.1	0.6	4.79	0	10.4
5 taps, 2nd	0.9	0.6	4.79	2	7.5
6 taps, 1st	3.1	0.6	6.58	2	9.8
6 taps, 2nd	4.8	0.8	6.58	2	9.7
7 taps, 1st	8.5	3.7	8.53	2	6.1
7 taps, 2nd	8.5	3.7	8.53	2	13.3

[a] The CBT have nine blocks arranged irregularly on a 23 × 28 cm board, Reversal corresponds to changing from clockwise to counter-clockwise and the other way around in CBT, not counting forward:backward

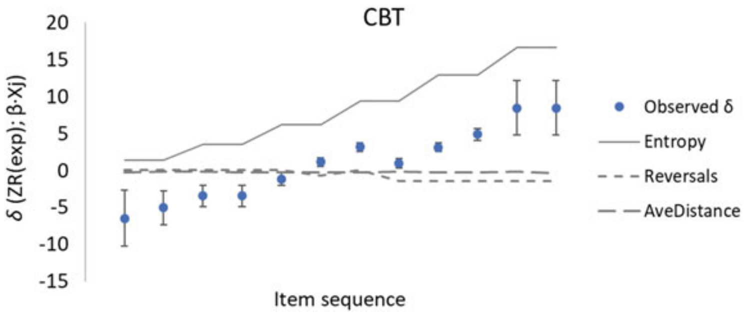


Fig. 10.1 Predicted contributions, $\Delta\delta$, to task difficulty from the three explanatory variables *Entropy, Reversals* and *Average distance* for the CBT. Item sequence corresponds to Table 10.1

Here the term *Entropy* is evaluated using the Brillouin [7] expression (Eq. 10.4) as a function of solely the number of blocks tapped. (That is, other effects discussed above, such as reduced entropy associated with symmetrical patterns and serial position effects such as primacy and recency, were found in the present studies to be negligibly small.) The contributions – or in fact the lack of contributions – from other explanatory variables are also illustrated in Fig. 10.1 where it is evident how *Entropy* increases with observed memory task difficulty while *Reversals* and *Average distances* only provide some noise around 0. The CSE have also larger measurement uncertainties for β-coefficients (Eq. 10.2) for *Reversals* and *Average distance* compared to *Entropy*.

As mentioned in above, a second but connected PCA is commonly performed to examine the unidimensionality of item attributes by examination of the residuals of the logistic regression of the Rasch measurement model to the observed outcomes. In the present case of CBT, this second PCA indicated only a weak additional dimension (where the first PC is the primary Rasch attribute) as a 1st contrast contributing as little as 9.9% unexplained variance. The two PCAs thus yield results here which are connected as expected: one dominant explanatory variable revealed in the first step of formulating a CSE with PCR is likely to be accompanied by indications of one single dimension in a Rasch residual PCA.

By regressing the observed memory task difficulty values, δ, against corresponding estimated zR from the CSE, the R^2 index indicates high accuracy of the prediction (Pearson coefficient $r = \sqrt{R^2} = 0.98$) (Fig. 10.2) and the observed memory task difficulty values, δ, were found to lie within the corridor of predicted uncertainties. i.e., $zR \pm UzR$ (Fig. 10.3).

The CSE for CBT illustrated here is very similar (within uncertainties) to a recently derived CSE for memory task difficulty in KCT [40, 52]:

$$zR_j = -9(5) + 2(1) \times Entropy_j + 0.8(1.4) \times Reversals_j + 0.7(2.9) \times AveDistance_j$$

At the same time, in earlier formulated CSEs for KCT, entropy was not considered, and distance and number of taps were confounded, making comparisons difficult.

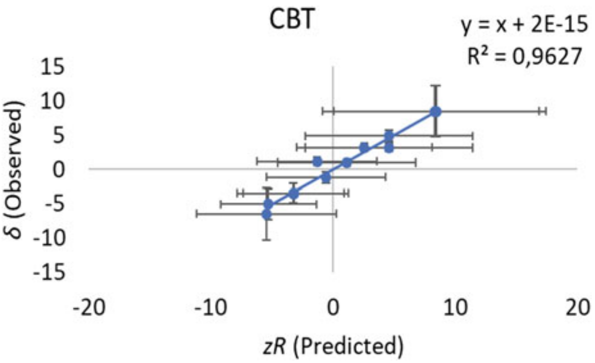


Fig. 10.2 Linear regression of the observed CBT memory task difficulty, δ, against the CSE predicted *zR* for the CBT. Uncertainties coverage factor $k = 2$. Item sequence corresponds to Table 10.1

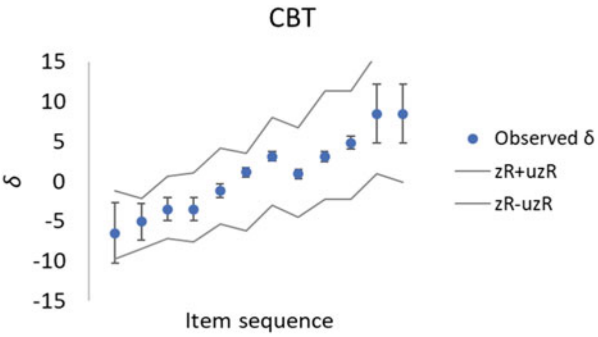


Fig. 10.3 Dots with uncertainty intervals shows the CBT observed memory task difficulty, δ and corridors of modelled uncertainties shows *zR*$\pm$*UzR* (grey lines) for the predicted *zR* values, coverage factor $k = 2$. Item sequence corresponds to Table 10.1

10.5.3.2 Case 2: Digit Recall

The memory task difficulty values, δ, ranged from 8.5 ±3.7 to 6.8 ± 2.2 logits in DTS. As for the CBT, the shortest tapping sequences were easier than the longer, but, although larger measurement uncertainties, the two longest sequences differ quite a lot in their observed memory task difficulty, δ. This is explained by 9 of 21 subjects remembered the 1st 8-digit sequence whilst only 2 of 21 remembered the 2nd 8-digit sequence. The item reliability was 0.96 and fit statistics were satisfactory, although, the two shortest sequences were classified as *minimum measures* due to all 86 subjects took the test passed. A summary of the memory task difficulty values, δ, and corresponding explanatory variables used for developing a CSE for DST are presented in Table 10.2.

The CSE for DST is very similar to the above illustrated CSE for CBT with *Entropy* as the dominating term and larger measurement uncertainties for β-coefficient (Eq. 10.2) for *Reversals* and *Average distance* compared to *Entropy* (Fig. 10.4):

$$zR_j = -10(3) + 1.5(3) \times Entropy_j + 0.2(4) \times Reversals_j - 0.2(1,5) \times AveDistance_j$$

Table 10.2 The Rasch estimates, δ_j, for each item, *j*, and its explanatory variables in DST

Number sequence	Observed memory task difficulty, δ	U (δ)	Entropy	Reversals	Average distance
3 digits, 1st	−8.5	3.7	1.79	1	3.7
3 digits, 2nd	−8.5	3.7	1.79	1	2.0
4 digits, 1st	−5.4	1.2	3.18	1	2.5
4 digits, 2nd	−5.4	1.2	3.18	2	3.5
5 digits, 1st	−3.4	0.8	4.79	2	3.2
5 digits, 2nd	−2.0	0.7	4.79	3	3.4
6 digits, 1st	0.8	0.6	6.58	3	4.3
6 digits, 2nd	−0.1	0.6	6.58	4	2.7
7 digits, 1st	3.3	0.9	8.53	2	2.1
7 digits, 2nd	2.3	0.8	8.53	3	3.6
8 digits, 1st	3.3	1.1	10.60	5	4.4
8 digits, 2nd	6.8	2.2	10.60	6	4.3

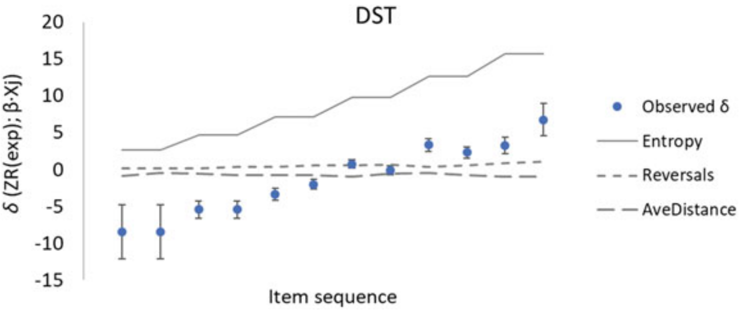


Fig. 10.4 Predicted contributions, Δ*δ*, to task difficulty from the three explanatory variables *Entropy, Reversals* and *Average distance* for the DST. Item sequence corresponds to Table 10.2

This is further illustrated in Fig. 10.5. Figure 10.5 shows the regressed observed memory task difficulty values, *δ,* against corresponding estimated $\underline{zR}$ from the CSE, the R^2 index indicates high accuracy of the prediction (Pearson coefficient r $= \sqrt{R^2} = 0.98$) and Fig. 10.6 shows the observed memory task difficulty values, *δ,* lying within the corridor of predicted uncertainties. i.e., *zR* ± *UzR*.

The DST sequences used here did not include any repeated digits, which however do occur, for example, in a 9-digit sequence in another version of DST. That data were retrieved from GBG MCI study [71], which included 268 individual tests involving 213 MCI patients (79%) and 55 HC patients (21%). The same procedure for transforming raw data into δ parameters for task difficulty was applied. In the GBG MCI study [71], three digits are repeated twice each. This has implications for the entropy contribution from the replicated digits [40, 44]. The sequence exemplified has higher order and convey less information compared to a digit sequence with similar number of digits but no duplicates. This is illustrated to the right in Fig. 10.7;

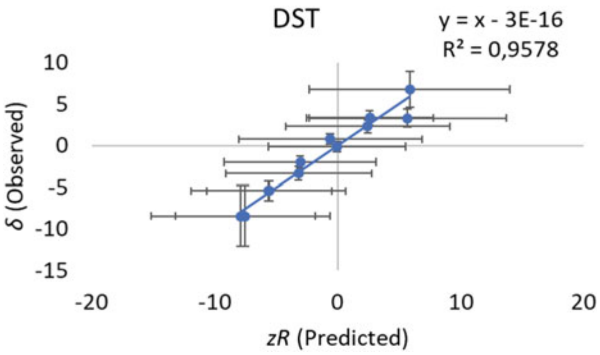


Fig. 10.5 Linear regression of the observed DST memory task difficulty, δ, against the CSE predicted *zR* for the DST. Uncertainties coverage factor $k = 2$. Item sequence corresponds to Table 10.2

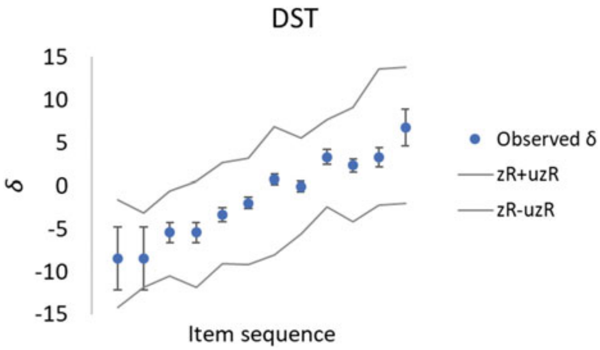


Fig. 10.6 Dots with uncertainty intervals shows the DST observed memory task difficulty, δ and corridors of modelled uncertainties shows $zR \underline{+} UzR$ (grey lines) for the predicted *zR* values, coverage factor $k = 2$. Item sequence corresponds to Table 10.2

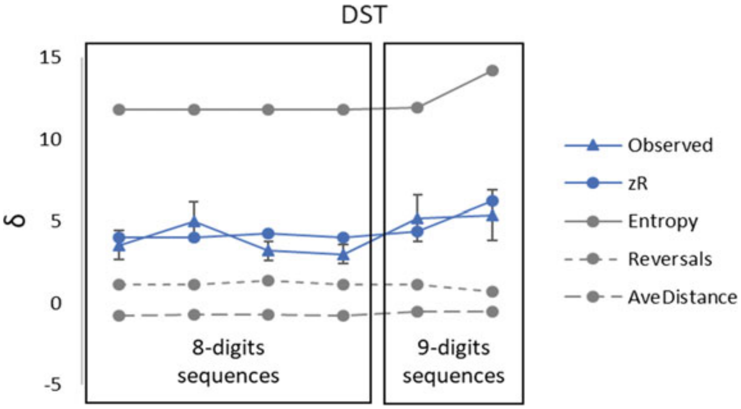


Fig. 10.7 Blue triangles for observed DST [GBG MCI study] memory task difficulty, δ, and blue dots for predicted *zR* task difficulty for 8-digits and 9-digits sequences, respectively, and their corresponding explanatory variables

the prediction of that item is easier than another other 9-digit sequence without duplicates. Likewise, as also illustrated to the left in Fig. 10.7, that item has a corresponding entropy contribution and predicted task difficulty which turn out to be comparable with those properties of the 8-digits items studied here, thus indicating a degree of equivalence between items of different tests. Consequently, this also complements the theoretical justification for considering an information theoretical approach including *Entropy* as an explanatory variable when explaining memory task difficulty.

10.5.3.3 Case 3: Taps and Digits Combined

The similarities in the administration of CBT and DST as well as their CSEs, respectively, invite to combine the items into one set of memory tasks. The item reliability was 0.93 and the memory task difficulty values, δ, ranged from −6.4 ± 3.7 to 6.8 ± 3.7 logits (and person ability values, θ, to be used in next section ranged from −4.0 ± 2.1 to 3.2 ± 1.5 logits). Three items showed *Maximum measures* (all from CBT) and two items showed *Minimum measures* (both from DST).

The main advantage of combining block and digit items in an extended CSE is that, with more information, uncertainties for assessing the persons' memory abilities reduces. This is illustrated in Fig. 10.8 where uncertainties, U(θ), for each person's memory ability are reduced approximately with 0.5 when comparing with U(θ) derived from only CBT.

Moreover, scaling CBT and DST together enables not only a hierarchical ordering of tasks, but also numerical values with which to compare the different recall sequences. As shown in Fig. 10.9, the two easiest DST memory tasks, i.e., 3-digit sequences are approximately 2 logits easier than corresponding 3-tapping sequences for CBT. Similarly, CBT memory tasks are somewhat more challenging than DST, even though the two sets of sequence tests have roughly the same number of entities i.e., lengths of sequence. However, with the same length, and without duplicates, predicted task difficulty zR for item j would be the same if entropy were the only explanatory variable, but those 3-taps/digits sequences differ in the number of reversals and the average distance (Tables 10.1 and 10.2).

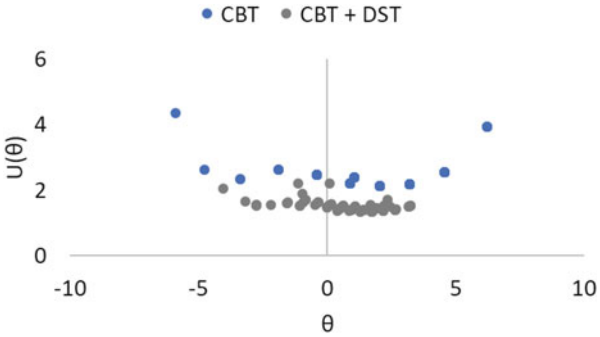


Fig. 10.8 Measurement uncertainties (U(θ), y-axis) for each person's memory ability from CBT in blue and from the combined CBT and DST in grey. At x-axis, from left to right least able persons to most able persons. Measurement uncertainties coverage factor $k = 2$

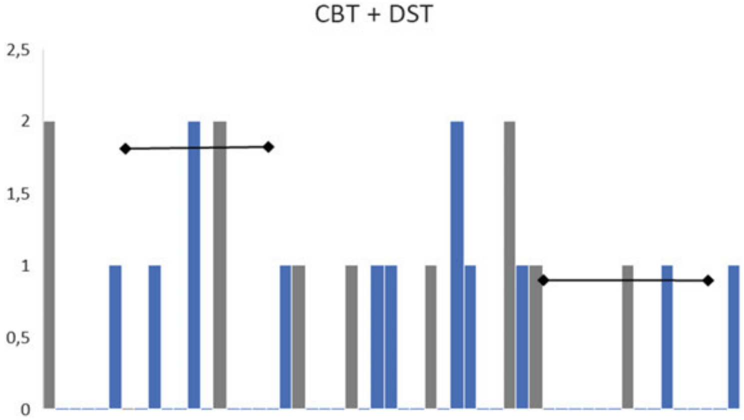


Fig. 10.9 Histogram of memory task difficulty for CBT in blue and DST in grey scaled together from left to right easiest tasks to most challenging tasks. Item order corresponds to Table 10.1 for CBT and Table 10.2 for DST. Error bars indicate measurement uncertainties coverage factor $k = 2$

10.5.4 Explaining Person Memory Ability

The previous section deal with a description of the objects itself, i.e., explaining memory tasks difficulties. CSEs or construct theories for task difficulty are the most common way of talking about the measurement validity. However, scores typically determined by identifying person locations along a single proficiency continuum do not in themselves naturally provide diagnostic information [15]. So, to recap, only when synthetic constructs created from theory enable the consistent and reliable prediction of both memory item (difficulty) and person memory (ability) location calibrations, can we claim to understand our memory measurements. Thus, the following section will provide an example on how a CSE for person memory ability might look like.

For the purpose of this demonstration the attribute of interest, ***Y***, were person memory abilities estimated based on the combination of CBT and DST assessments. The ‘something’ that causes variation in the attribute of interest are variables that can be used to explain why some persons have better memory abilities than others such as well-known AD-related biomarkers:

- cortical thickness (mm^3)
- left hippocampus volume (mm^3; normalized to each test person’s intercranial volume)
- neurofilament light (*NfL*) (pg/ml)
- amyloid peptides 1-42/1-40 ratio (*Aβ42/40*)
- total tau proteins (*Tau*) (pg/ml)

where the latter three were measured in plasma.

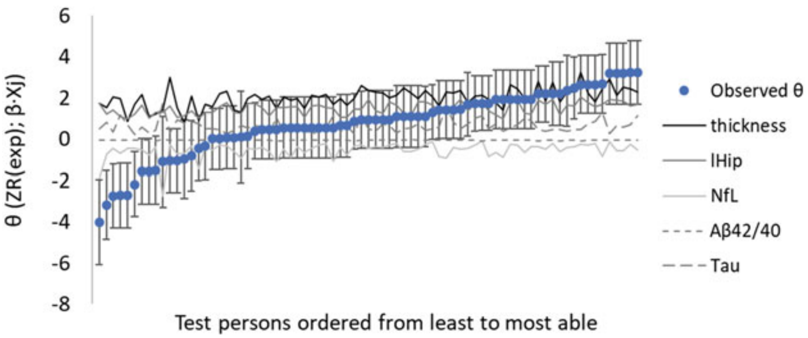


Fig. 10.10 Predicted contributions, Δθ, to person ability from the five explanatory variables *Thickness, left Hippocampus, NfL, Aβ42/40* and *Tau*

CSEs for Person Memory Ability Based on Biomarkers

The person memory ability values, θ, ranged from −4.0 ± 2.1 to 3.2 ± 1.5 logits with a person reliability of 0.70. To correspond to memory task difficulty, from easier tasks to more challenging tasks, a lower value of person memory ability indicates a less able person, while a higher value of person memory ability indicates a more able person.

The CSE yielded from the PCR where the five biomarkers were used to explain person memory ability showed a less clear contribution from a single biomarker compared to entropy as explanatory variable for memory task difficulty:

$$zR_i = -2(3) + 0.24(2) \times thickness^3 + 1(2) \times lHip - 0.03(3) \times NfL - 0.2(4.9) \times A\beta 42/40 + 0.4(4) \times Tau$$

As shown in the CSE as well as in Fig. 10.10; *NfL* has a negative contribution, while *cortical thickness* and *hippocampus* volume exhibits a positive contribution. This corresponds to the accepted clinical interpretation of disease progress where higher levels of *NfL* are seen in patients with AD compared to more able persons within the AD spectrum and brain volume is decreasing as the disease progresses. However, there were negligible contribution from *Aβ42/40* and *Tau* to explain person memory ability and the *U(β)* shown for each explanatory variable in the CSE are larger than their coefficient itself for Aβ42/40 and similar for *Tau*, while the three others have minor uncertainties.

The R^2 index indicates lower accuracy of the prediction compared to the CSEs for memory task difficulty presented above (Pearson coefficient $r = \sqrt{R^2} = 0.57$). At the same time the regression based on multivariable express higher accuracy than any of the individual biomarkers in univariate correlations (Person coefficient r ranging from 0.04 (*Tau*) to 0.51 (*Thickness*). Even though *Thickness* has close to similar Pearson coefficient as the CSE, the relative uncertainty for the β-coefficient is greatly reduced within the CSE compared to the univariate correlation (*Thickness* β-coefficient and *U*(β) in CSE/multivariate is 0.24 ± 0.02 compared to univariate 0.4 ± 0.2).

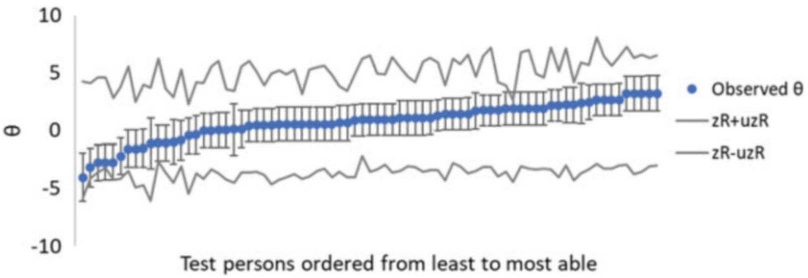


Fig. 10.11 Dots with uncertainty intervals shows the observed person memory ability, θ, and corridors of modelled uncertainties shows *zR*±*UzR* (grey lines) for the predicted *zR* values, coverage factor $k = 2$

Figure 10.11 illustrates a corridor of the *zR*+*UzR* around the U(θ) for each person's memory ability, which at some parts is considerably wider than the U(θ), probably because there are additional components of variation not yet included in the CSE model.

10.6 Limitations and Implications in Interpreting CSEs

Only a fully comprehensive multivariate and PCR study including all potentially important explanatory variables will give a true picture. A fully comprehensive CSE for either memory task difficulty or person memory ability are not yet evident in the cases illustrated- they all have their limitations. In their critical review of what they termed "decomposing item difficulties", Green and Smith [21] mentioned a number of potential general limitations in various multivariate and regression approaches – such as "effects of sample size, collinearity, a measurement disturbance, and multidimensionality on the estimation of component difficulties" These limitations are still valid today when forming CSEs for task difficulty, and also apply equally to person ability estimation as well.

With regards to the effects of sample size, it is well-known that large well-targeted samples provide more information about each item, and consequently, as the sample size increases, the measurement uncertainties for estimates of task difficulty are reduced [76]. And similarly, the other way around, with increased test length (i.e., number of items), measurement uncertainties are reduced for estimates of person ability, as illustrated in Fig. 10.8. However, as also mentioned earlier in this Chapter, the measurement uncertainties for each memory task difficulty, *u*(δ), and for each person's memory ability, *u*(θ), propagate through the PCR. In turn the u(δ) and u(θ), respectively, will have implications for *U(*β*)* and *UzR* together with uncertainties in the fit itself, which brings us to the issue of collinearity and measurement disturbance. As is evident from Figs. 10.3, 10.6 and 10.11, the "corridor" of modelled uncertainties, *UzR,* is (much) wider than the observed memory task difficulty, *δ*, and person memory ability, θ, with its measurement uncertainties, respectively. This we interpret as indicating that there are sources of dispersion when making the multivariate regression which are not accounted for in terms only of measurement uncertainty. Some such dispersion could arise from one or more additional explanatory variables of variation associated with unrecognized explanatory variables but could also be signs of collinearity and other disturbances and warrants further investigation. Examples of additional explanatory variables for person memory ability could be the brain entropy and connectivity, as well as, but not limited to biomarkers, the person's ability to pay attention to the task to be performed and his or her executive functions.

One way to reduce the limitations stemming from collinearity is by implementing the PCA in the first step of developing the CSE. With the PCA, the main components of variation (found by "rotating" in the explanatory-variable space from the experimental dimensions to the PC dimensions) can be obtained. Nevertheless, before that, one needs to consider each explanatory variable, **X** in terms of our understanding. For instance, to avoid collinearity, we have chosen to only include the left hippocampus volume, and not also the right hippocampus volume, even though both show univariate Pearson coefficients of approximately 0.45 with person memory ability. Left and right hippocampus volumes showed a Pearson coefficients $r = \sqrt{R^2} = 0.90$, and in turn, adding both did not provide any additional predictive value in the CSE.

We performed an initial study of multidimensionality above, for instance in the discussion about task difficulty for the tapping test CBT where the PCA of item residuals indicated one dimension, and corresponding one dominant explanatory variable. As mentioned by Green and Smith [21], unidimensionality is an assumption of the Rasch model and a straightforward use of regression will be of only marginal value unless the items form a cohesive set, that is, that the same underlying variable explains response to every item in that set. However, in general terms, as mentioned by Green and Smith [21], the less well-defined or well-constructed items are, e.g., by including non-fitting items or appearance of multidimensionality, the less likely it is that the appropriate model will be identified. This brings us back to the significance of proper construct modelling [75] including a substantive theory. In our case, there is however one possible limitation in terms of the subjects included and the choice of AD related biomarkers as explanatory variables **X**. Even though the memory abilities are ordered as expected – AD patients having lowest abilities and health controls having highest ability – it is not necessary that all persons' memory abilities only can be explained in terms of AD related biomarkers, e.g., a person with MCI might not progress to AD as it could also be other dementia characterized by other biomarkers.

Despite those limitations, the benefits are more important, e.g., when contrasting univariate and multivariate correlation studies. A main disadvantage of univariate fits is that the observed correlation between the attribute and a chosen explanatory variable might be explained by another covariate not considered. Where multivariate correlation studies can be reliably performed, lower measurement uncertainties than univariate fits are to be expected which is of great significance in the present field of memory measurements, and beyond.

10.7 Chapter Summary, Strengths and Future Recommendations

This chapter has focused on two elements to ensure understanding, measuring and quality-assuring constructs with examples of memory measurements; explicit methods for testing theories of the measurement mechanism and establishment of metrological standards; and substantive theories explaining the constructs themselves. We have demonstrated advancement in both those elements gained by adopting a measurement system approach, including modelling *Man as a Measurement Instrument*, exploring the principal components in explanatory variables for *Construct Specification Equations* (CSE) and introducing the concept of entropy as an explanatory variable.

Especially building on entropy, as described in thermodynamics and information theory, further enables more fit-for-purpose and valid memory measurements. The significance of a CSE for memory task difficulty in memory measurements is that it can facilitate establishing objective and scalable metrological units through the generation of certified reference 'materials' for traceability. Moreover, formulation of CSEs based on entropy for memory task difficulty in turn opens opportunities of formulating new, less onerous but more sensitive and representative tests largely by 'cherry-picking the best' items from existing batteries, such as exemplified with CBT and DST. Formulation of CSEs for person ability can be a means of providing diagnostic information to enhance clinical decisions and targeted interventions, although, there are probably additional components of variation (e.g., entropy-based measures of brain connectivity) not yet included in the CSE model to further improve the understanding what causes lower or higher memory ability.

Acknowledgments Some of the data used in the examples in this chapter has been collected as a part of the 15HLT04 NeuroMET project which received funding from the EMPIR programme co-financed by the Participating States and from the European Union's Horizon 2020 research and innovation programme. We would therefore offer our special thanks to partners at Charité University Hospital for patent recruitment and memory test data as well as Physikalisk Technisches Bundesanstalt (PTB), CHU Montpellier and LGC UK National Measurement Laboratory for analyses of brain volumes and biomarkers. We would also like to send our thanks to the Gothenburg MCI study for letting us use Digit Span Test (DST) data from their cohort.

Appendix: PCR Algorithms

The various terms on both sides of each PCR equation [section 1.4] were evaluated in this work using several modules (Table 10.3) of the program PTC® MathCad®, which is a high-level, matrix-based language with which text and illustrations can be combined with data handling (such as input/output), calculations and graphs in an active way. This has the advantage of providing immediate and explicit feedback when designing and formulating new programming routines as well as clear documentation for communicating with third parties and for future retrievable archiving.

Table 10.3 Mathematical expressions in PCR and corresponding PTC® MathCad® modules

Mathematical expression	PTC® MathCad® module	PTC® MathCad® Description
PCA		
$Cov(\mathbf{X})$	$Covar(\mathbf{X})$	Returns the covariance of matrix X. Implements methods and algorithms described in the books mentioned in the PTC® MathCad® *Signal Processing Bibliography*
λ	λ ≔ reverse(sort (eigenvals ($Covar(\mathbf{X})$)))	
	reverse(A)	Reverses the order of elements in a vector, or the order of rows in a matrix A. Uses a heapsort algorithm (Press, *et. al,* Numerical Recipes).
	sort(v)	Returns a vector with the values from v sorted in ascending order. Uses a heapsort algorithm (Press, *et. al*, Numerical Recipes).
	eigenvals (M)	Returns a vector whose elements are the eigenvalues of M. Uses the Intel Basic Linear Algebra Subprograms (BLAS)/Linear Algebra Package (LAPACK) libraries.
$\mathbf{p}_n$	$P^{\langle n\rangle}$ ≔ eigenvec ($Covar(\mathbf{X}), \lambda_n$)	
	eigenvec($\mathbf{M}$, z)	Returns a single normalized eigenvector associated with eigenvalue z of M. The eigenvec function uses an inverse iteration algorithm from the Intel Basic Linear Algebra Subprograms (BLAS)/Linear Algebra Package (LAPACK) libraries.
Regression, Eq. 10.3	$polyfitc(X1, ZR, m)$ $X1 = \mathbf{X}'$; $ZR = \mathbf{Y}$; $m = 1$; fit order	Returns the regression coefficients for a multivariate polynomial regression surface fitting the results recorded in matrix Y to the data found in matrix $\mathbf{X}'$. Column 1 of *polyfitc*: Regression coefficient for each term
$\mathbf{u}\left(\widehat{\mathbf{C}}\right)$		Column 2 of *polyfitc*: Standard error for the regression coefficient

References

1. N.D. Adroher, A. Tennant, Supporting construct validity of the Evaluation of Daily Activity Questionnaire using Linear Logistic Test Models. Qual. Life Res. **28**(6), 1627–1639 (2019). https://doi.org/10.1007/s11136-019-02146-4
2. D. Andrich, Implications and applications of modern test theory in the context of outcomes based education. Stud. Educ. Eval. **28**, 103–121 (2002)
3. ASTM, Standard Guide for Measurement Systems Analysis (MSA) E2782. (2012), https://doi.org/10.1520/E2782-11
4. C. Boake, From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. J. Clin. Exp. Neuropsychol. **24**(3), 383–405 (2002). https://doi.org/10.1076/jcen.24.3.383.981
5. D. Borsboom, *Measuring the Mind. Conceptual Issues in Contemporary Psychometrics* (Cambridge University Press, Cambridge, 2005)
6. D. Borsboom, G.J. Mellenbergh, J. van Heerden, The theoretical status of latent variables. Psychol. Rev. **110**(2), 203–219 (2003). https://doi.org/10.1037/0033-295x.110.2.203
7. L. Brillouin, *Science and Information Theory* (Academic, New York, 1962)
8. R.M. Busch, K. Farrell, K. Lisdahl-Medina, R. Krikorian, Corsi Block-Tapping task performance as a function of path configuration. J. Clin. Exp. Neuropsychol. **27**(1), 127–134 (2005). https://doi.org/10.1080/138033990513681
9. S.J. Cano, L.R. Pendrill, J. Melin, W.P. Fisher, Towards consensus measurement standards for patient-centered outcomes. Measurement **141**, 62–69 (2019). https://doi.org/10.1016/j.measurement.2019.03.056
10. L. Carnot, *Principes fondamentaux de l'équilibre et du mouvement;* Par L.N.M. C*arnot*. De l'imprimerie de Crapelet. A Paris, chez Deterville, libraire, rue du Battoir, no 16, quartier S. André-des-Arcs. An XI-1803 (1803). https://play.google.com/books/reader?id=kslJAAAAcAAJ&pg=GBS.PA254&hl=en_GB
11. K.B. Christensen, Latent covariates in generalized linear models: A Rasch model approach, in *Advances in statistical methods for the health sciences: Applications to Cancer and AIDS Studies, genome sequence analysis, and survival analysis*, ed. by J.-L. Auget, N. Balakrishnan, M. Mesbah, G. Molenberghs, (Birkhäuser, Boston, 2007), pp. 95–108
12. R. Clausius, Ueber die bewegende Kraft der Wärme und die Gesetze, welche sich daraus für die Wärmelehre selbst ableiten lassen. Ann. Phys. **79**(4), 368–397, 500–524 (1850). https://doi.org/10.1002/andp.18501550403
13. P.M. Corsi, *Human Memory and the Medial Temporal Region of the Brain* (34) (ProQuest Information & Learning, US, 1973)
14. P. De Boeck, M. Wilson (eds.), *Explanatory item response models: A generalized linear and nonlinear approach*, Statistics for social and behavioral sciences (Springer, New York, 2004)
15. J. de la Torre, N. Minchen, Cognitively diagnostic assessments and the cognitive diagnosis model framework. Psicología Educativa **20**(2), 89–97 (2014)
16. R. Dybkaer, *An Ontology on Property for Physical, Chemical, and Biological Systems* (2009). ISBN 978-87-990010-1-9. https://ontology.iupac.org/
17. EMPIR 15HLT04, *Innovative measurements for improved diagnosis and management of neurodegenerative diseases*. Retrieved from https://www.lgcgroup.com/our-programmes/empir-neuromet/neuromet/
18. G. Fischer, The linear logistic test model as an instrument in educational research. Acta Psychol. **37**, 359–374 (1973)
19. W.P. Fisher Jr., Food for thought from Carnot. Popul. Meas. **4**(1), 13–14 (2002) https://rasch.org/pm/pm4.pdf
20. W.P. Fisher, M. Wilson, An online platform for sociocognitive metrology: The BEAR Assessment System Software. Meas. Sci. Technol. **31**(3) (2019)
21. K. Green, R. Smith, A comparison of two methods of decomposing item difficulties. J. Educ. Stat. **12**, 369–381 (1987)

22. J. Hobart, S. Cano, H. Posner, O. Selnes, Y. Stern, R. Thomas, J. Zajicek, Putting the Alzheimer's cognitive test to the test I: Traditional psychometric methods. Alzheimers Dement. **9**(1 Suppl), S4–S9 (2013). https://doi.org/10.1016/j.jalz.2012.08.005
23. L.F. Hughes, K. Perkins, B.D. Wright, H. Westrick, Using a Rasch scale to characterize the clinical features of patients with a clinical diagnosis of uncertain, probable, or possible Alzheimer disease at intake. J. Alzheimers Dis. **5**(5), 367–373 (2003). https://doi.org/10.3233/jad-2003-5503
24. S.M. Humphry, The role of the unit in physics and psychometrics. Meas. Interdiscip. Res. Perspect. **9**(1), 1–24 (2011)
25. ILAC G9:2005, *Guidelines for the Selection and Use of Reference Materials* (International Laboratory Accreditation Committee, 2005)
26. I. Imbo, A. Szmalec, A. Vandierendonck, The role of structure in age-related increases in visuo-spatial working memory span. Psychol. Belgica **49**, 275–291 (2009)
27. ISO Guide 33:2015, *Reference Materials – Good Practice in Using Reference Materials* (International Organization for Standardization, Pretoria, 2015)
28. G. Iverson, R. Luce, *The Representational Measurement Approach to Psychophysical and Judgmental Problems, in Measurement, Judgment, and Decision Making* (Academic, Cambridge, 1998)
29. C.R. Jack Jr., D.A. Bennett, K. Blennow, M.C. Carrillo, B. Dunn, S.B. Haeberlein, et al., NIA-AA research framework: Toward a biological definition of Alzheimer's disease. Alzheimers Dement. **14**(4), 535–562 (2018). https://doi.org/10.1016/j.jalz.2018.02.018
30. A. Kamata, *One-Parameter Hierarchical Generalized Linear Logistic Model: An Application of HGLM to IRT*. Paper presented at the annual meeting of American Educational Research Association, San Diego, CA, April (1998)
31. W.T. Kelvin, An account of Carnot's theory of the motive power of heat – With numerical results deduced from Regnault's experiments on steam. Trans. Edinb. R. Soc. **XVI** (1849)
32. H. Knox, A scale, based on the work at Ellis Island, for estimating mental defect. J. Am. Med. Assoc. **LXII**(10), 741–747 (1914)
33. A. Kyngdon, The Rasch Model from the perspective of the representational theory of measurement. Theory Psychol. **18**(1), 89–110 (2008)
34. S.L. Latimer, Using the Linear Logistic Test Model to investigate a discourse-based model of reading comprehension. Educ. Res. Persp. **9**(1), 73–94 (1982)
35. J.M. Linacre, Structure in Rasch residuals: Why principal components analysis (PCA)? Rasch Meas. Trans. **12**(2), 636 (1998) https://www.rasch.org/rmt/rmt122m.htm
36. J.M. Linacre, B.D. Wright, The length of a logit. Rasch Meas. Trans. **3**(2), 54–55 (1989)
37. R.D. Luce, J.W. Tukey, Simultaneous conjoint measurement: A new type of fundamental measurement. J. Math. Psychol. **1**, 1–27 (1964)
38. J. Mattila, J. Koikkalainen, A. Virkki, A. Simonsen, M. van Gils, G. Waldemar, et al., A disease state fingerprint for evaluation of Alzheimer's disease. J. Alzheimers Dis. **27**, 163–176 (2011)
39. G.M. McKhann, D.S. Knopman, H. Chertkow, B.T. Hyman, C.R. Jack Jr., C.H. Kawas, et al., The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. **7**(3), 263–269 (2011). https://doi.org/10.1016/j.jalz.2011.03.005
40. J. Melin, L.R. Pendrill, S.J. Cano, EMPIR NeuroMET 15HLT04 consortium, Towards patient-centred cognition metrics. J. Phys. Conf. Ser. (2019). https://doi.org/10.1088/1742-6596/1379/1/012029
41. J. Melin, W.P. Fisher, L.R. Pendrill, *A Hierarchy of Construct Theories: Their Focus and Manifestations*. Paper presented at the International Objective Measurement Workshop (IOMW) Conference, Berkeley, CA (2020, April). https://www.iomw.org/
42. J. Melin, S.J. Cano, A. Flöel, L. Göschel, L.R. Pendrill, Construct specification equations: 'Recipes' for certified materials in cognitive measurement. Meas. Sens. **18**, 100290 (2021a)

43. J. Melin, S.J. Cano, A. Regnault, L.R. Pendrill, *Neuropsychological Assessments – Word Learning List Memory Tests and Diagnostic Potential of Serial Position Effects* (CIM, Lyon, 2021b)
44. J. Melin, S. Cano, L. Pendrill, The role of entropy in construct specification equations (CSE) to improve the validity of memory tests. Entropy **23**(2), 212 (2021c). https://doi.org/10.3390/e23020212
45. J. Melin, S.J. Cano, L. Göschel, A. Fillmer, S. Lehmann, C. Hirtz, A. Flöel, L.R. Pendrill, Metrological references for person ability in memory test. Meas. Sens. **18**, 100289 (2021d)
46. S.S. Menon, K. Krishnamurthy, A study of brain neuronal and functional complexities estimated using multiscale entropy in healthy young adults. Entropy **21**(10), 995 (2019)
47. J. Michell, *Measurement in Psychology: A Critical History of a Methodological Concept* (Cambridge University Press, Cambridge, 1999) ISBN-10: 0521621208
48. S.B. Nutley, S. Soderqvist, S. Bryde, K. Humphreys, T. Klingberg, Measuring working memory capacity with greater precision in the lower capacity ranges. Dev. Neuropsychol. **35**(1), 81–95 (2010). https://doi.org/10.1080/87565640903325741
49. L.R. Pendrill, Risk assessment and decision-making risk assessment and decision-making, in *Theory and Methods of Measurements with Persons*, ed. by B. Berglund (Stockholm, SE), G.B. Rossi (Genoa, IT), J. Townsend (Bloomington, IN), L.R. Pendrill (Borås, SE) (Psychology Press, Taylor & Francis, 2010). ISBN: 978-1-84872-939-1
50. L. Pendrill, Man as a measurement instrument. NCSLI Meas. **9**(4), 24–35 (2014). https://doi.org/10.1080/19315775.2014.11721702
51. L.R. Pendrill, Assuring measurement quality in person-centred healthcare. *Measurement Science and Technology, 29*(3), 034003 (2018). https://doi.org/10.1088/1361-6501/aa9cd2
52. L. Pendrill, *Quality Assured Measurement, Unification Across Social and Physical Sciences* (Springer, 2019)
53. L. Pendrill, J. Melin, Assuring measurement quality in person-centred care, in *Person Centered Outcome Metrology*, (Springer, 2020)
54. B. Prien, How to predetermine the difficulty of items of examinations and standardized tests. Stud. Educ. Eval. **15**(3), 309–317 (1989). https://doi.org/10.1016/0191-491X(89)90012-6
55. L.L. Raket, Statistical disease progression modeling in Alzheimer disease. Front. Big Data (2020). https://doi.org/10.3389/fdata.2020.00024
56. G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests* (Danish Institute for Education Research, Copenhagen, 1960) (Expanded edition (1980) with foreword and afterword by B.D. Wright, University of Chicago Press, 1980. Reprinted Chicago: MESA Press, 1993.
57. G. Rasch, *On General Laws and the Meaning of Measurement in Psychology*. Paper presented at the Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Medicine, Berkeley, Calif (1961)
58. G. Rasch, *On Objectivity and Specificity of the Probabilistic Basis for Testing* (n.d.). Retrieved from https://www.rasch.org/memo196x.pdf
59. G.B. Rossi, *Measurement and Probability [Elektronisk resurs] A Probabilistic Theory of Measurement with Applications* (Springer, 2014)
60. T. Salzberger, Does the Rasch Model convert an ordinal scale into an interval scale? Rasch Meas. Trans. **24**(2), 1273–1275 (2010)
61. M.M. Schnore, J.T. Partington, Immediate memory for visual patterns: Symmetry and amount of information. Psychon. Sci. **8**, 421–422 (1967)
62. C.E. Shannon, W. Weaver, *The Mathematical Theory of Communication*, 2nd edn. (University of Illinois Press, Urbana, 1963)
63. A.J. Stenner, W.P. Fisher Jr., Metrological traceability in the social sciences: A model from reading measurement. J. Phys. Conf. Ser. **459**, 012025 (2013)
64. A.J. Stenner, M. Smith, Testing Construct theories. Percept. Mot. Skills **55**, 415–426 (1982)
65. A.J. Stenner, M. Smith, D. Burdick, Towards a theory of construct definition. J. Educ. Meas. **20**(4), 305–316 (1983)

66. A.J. Stenner, H. Burdick, E.E. Sandford, D.S. Burdick, How accurate are Lexile text measures? J. Appl. Meas. **7**(3), 307–322 (2006)
67. A.J. Stenner, W.P. Fisher Jr., M.H. Stone, D.S. Burdick, Causal Rasch models. Front. Psychol. **4**, 536 (2013). https://doi.org/10.3389/fpsyg.2013.00536
68. M. Stone, J. Stenner, *Substantive Theory and Constructive Measures* (iUniverse, 2018)
69. L. Tesio, Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J. Rehabil. Med. **35**, 105–115 (2003)
70. L.L. Thurstone, E.J. Chave, *The Measurement of Attitude: A Psychophysical Method and Some Experiments with a Scale for Measuring Attitude Toward the Church* (The University of Chicago Press, 1929). https://doi.org/10.1037/11574-000
71. A. Wallin, A. Nordlund, M. Jonsson, K. Lind, Å. Edman, M. Göthlin, et al., The Gothenburg MCI study: Design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. J. Cereb. Blood Flow Metab. **36**(1), 114–131 (2016). https://doi.org/10.1038/jcbfm.2015.147
72. Z. Wang, Y. Li, A.R. Childress, J.A. Detre, Brain entropy mapping esing fMRI. PLoS One **9**(3), e89948 (2014). https://doi.org/10.1371/journal.pone.0089948
73. D.J.J. Wang, K. Jann, C. Fan, Y. Qiao, Y.-F. Zang, H. Lu, Y. Yang, Neurophysiological basis of multi-scale entropy of brain complexity and its relationship with functional connectivity. Front. Neurosci. ***12***, 352–352 (2018). https://doi.org/10.3389/fnins.2018.00352
74. D. Wechsler, *Wechsler Adult Intelligence Scale* (Psychological Corp, San Antonio, 1955)
75. M. Wilson, *Constructing Measures: An Item Response Modeling Approach* (Erlbaum, Hoboken, 2005)
76. B. Wright, Error, variances and correlations. Rasch Meas. Trans. **5**(2), 147 (1991)
77. B. Wright, M. Stone, *Best Test Design: Rasch Measurement* (MESA Press, Chicago, 1979)

Chapter 11
Assuring Measurement Quality in Person-Centered Care

Leslie R. Pendrill and Jeanette Melin

Abstract The quality-assurance of measurement in person-centered care (PCC) – is introduced firstly by "bookending" the topic in the overall context of the quality assurance of the care itself. At the start the chapter we ask: What are the end-user objects and constructs of PCC – for instance as specified by the profession and in legislation? At the end: What decisions about PCC objects and constructs can be made and how reliable are they? Examples and illustrations from PCC have included (i) neuropsychological cases (dealt with in more detail in the accompanying chapter by Melin and Pendrill (Person centered outcome metrology. Springer, 2022)) and (ii) patient participation. In the two central sections of the chapter, assuring the quality of measurement in PCC has obliged consideration of how traditional metrological concepts – particularly metrological references for comparability via traceability and reliable estimates of uncertainty – need to be extended. In providing an overview of the benefits of combining Rasch measurement theory and quality assurance, the unique properties of Rasch Measurement Theory are exploited to the full. Replacing the instrument at the heart of a traditional measurement system with a human being provides a truly "person-centered" model of the metrology. This in turn enables a viable procedure to establish metrological references in fields such as PCC in the form of "recipes" analogous to certified reference materials or procedures in analytical chemistry and materials science. It also informs the measurement uncertainties which determine the final decisions about PCC taken at the end of the chapter.

Keywords Reference standards · Quality assured traceability · Uncertainty

L. R. Pendrill (✉) · J. Melin
Research Institutes of Sweden, Gothenburg, Sweden
e-mail: leslie.pendrill@ri.se

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3_11

11.1 Introducing Quality-Assurance of Measurement in Person-Centered Care

In person-centered care (PCC), the patient is first and foremost regarded as a person with reason, will, feeling and needs [24] with unique and holistic properties which need to be brought into a partnership with health care professionals (e.g., [24, 30, 44, 47, 59]). When quantifying symptoms and experiences in PCC, patient centered outcomes (PCOs) pertain to a patient's beliefs, opinions and needs in conjunction with a clinician's medical expertise and assessment. To capture faithfully these special aspects requires a special "person-centered" measurement process – typically with tests of performance, questionnaires and rating scales – alongside more traditional kinds of measurement in healthcare.

PCC is one of several applications where increased attention to quality assurance in healthcare is driving the discipline of quality-assured measurement – that is, metrology – as a topic of burgeoning and increasingly multidisciplinary interest. Quality assurance in a field such PCC needs conformity assessment [35] in order to provide

A. Confidence for the consumer (the patient) that requirements on products and services (care) are met
B. The producer and supplier (the health care organization) with useful tools to ensure product and service (care) quality
C. Help to regulators when ensuring that health, safety or environmental conditions are met.

Quality-assured measurement as a discipline has several centuries of history behind it, responding to quality-assurance needs in mainly technical domains such as trade and industrial production where physicists and engineers have led the field. At the start of the twenty-first Century, there is now a need and a challenge to formulate a unified view of metrology to address contemporary concerns and grand challenges, not only in physics and engineering but also in the social sciences [90]. PCC is one particular area of focus, [16] where some essential measurement aspects have been emphasized, for instance, by Walton et al. [94] and Cano et al. [17] and are considered further in the section 11.1.2 *Quality assurance in person-centered care. Design of experiments*.

Because metrology is not an end in itself, the start of this chapter – as well as at the end – will not deal directly with measurement, but rather with the objects – products, services, concepts… and their characteristics in PCC, particularly the patient centered outcomes (PCO). It is the latter which are the concern of the vast majority of people working in PCC, who then ask metrologists to measure them. Presenting the assurance of measurement quality in person-centered care takes the approach advocated more generally in another book, *Quality Assured Measurement* [75] in this Springer Series. The start and end of the chapter provide object-related 'bookends' – supporting a description of Quality-Assured Measurement which is the central issue and dealt with in the middle sections of this chapter.

Presenting measurement in relation to assessed "objects" (tasks, cohort individuals, clinicians), rather than as an end in itself, will allow measurements to be anchored in relevance and interest for third parties, which in the context of PCC can be all stakeholders: from patient, health care professionals, relations, health care organizations, to regulators, politicians, and decision-makers.

As it turns out, the approach also provides the key to a unified presentation about quality-assured measurement across Social and Physical Sciences where objects are probed by a *Human as a Measurement Instrument* in "person-centered" measurement processes and in Measurement System Analysis (MSA) [1, 7], as explained further in Pendrill [75] (2019 and the section 11.2.3 *A way forward for measuring PCOs: Human as a B: Measurement Instrument*).

11.1.1 Opening the Quality-Assurance Loop

Since neither production nor measurement processes are perfect, assuring the quality of the product or other entity (process, phenomenon, etc.) – in the present case, PCC as assessed in terms of PCOs – will involve efforts to keep within tolerable limits the unavoidable, real or apparent dispersion in the entity value. (Wherever possible, use will be made of the international vocabulary for conformity assessment in choice of terminology [38].

Production errors (addressed in the section 11.1.3 *A*: *Entity attribute description and specification*) and apparent errors arising from an imperfect measurement system (addressed in the section 11.2 *Benefits of combining Rasch Measurement Theory (RMT) and quality assurance*) can both be revealed by measurement, either of a series of items or repeated observations of a single item.

"Item" in the context of PCC refers to a particular example of product – for example, an instance of delivering a care service to an individual patient (see further in the section 11.1.2 *Quality assurance in person-centered care. Design of experiments*).

Limited knowledge – about production or about measurement – will lead, respectively, to uncertainties in both product error and measurement error (as described in the section 11.3.2 *Errors and uncertainties in PCOs*); each having associated risks of incorrect decisions of conformity (as described in the section 11.4 *Decision risks and uncertainty*).

Conformity assessment is a formal process with the specific aim of keeping product "on target". In many cases, both product and measurement processes will be subject to conformity assessment.

The overall concept of a quality-assurance loop was developed over 60 years ago in the context of industrial production, emphasizing that product quality is not only determined by actions at the point of production, but importantly at every step of the loop – from initially defining, designing to finally delivering product – in what should be an on-going 'dialogue' between consumer and supplier [21].

The rules and regulations embodying the demands of decision-makers, regulators and authorities, together with expert and end-user opinion and advice for each sector – such as in healthcare (section 11.1.2 *Quality assurance in person-centered care. Design of experiments*) – can be formulated in clear, unambiguous and harmonized ways in written standards and norms suitable for implementation when assessing quality. The quality-assurance loop has provided a framework for major series of international standards for quality assurance, particularly the famous ISO 9000 series which over the years have been adapted not only for industrial production but also healthcare, such as the European norm: EN 15224:2017, *Quality management systems – EN 9001:2015 for healthcare*, where clinical processes in health care services are the main focus. At the same time, quality-assured measurement informs the written standards to ensure the viability of the regulatory demands in terms of what is practically and realistically possible to measure as part of an overall quality infrastructure (figure 6.1 in [75]).

11.1.2 Quality Assurance in Person-Centered Care. Design of Experiments

The first step in the quality-assurance loop is to define “product” (in the widest sense). A valid and relevant formulation of the construct of interest is a key step.

Conformity assessment occurs in the framework of regulation where *clear definition of* “product” *is of course necessary*. As an example, according to current Swedish Health and Medical Services Act:

> The goal of health care is good health and care on equal terms for the entire population... This implies that health care activities must uphold a good hygienic standard; meet the patient's need for security, continuity and safety; build on respect for the patient's self-determination and integrity; promote good contacts between the patient and health care professionals, and be easily accessible. Moreover, the quality of the healthcare shall be systematically and continuously developed and assured. (HSL, 2017:30, chap. 3, General § 1 and chap. 5 Activities General Sections §§1–4) [31]

Several of these quality characteristics can also be found in the “ISO 9000-like” European norm for healthcare: EN 15224:2017 [25].

While not explicit in the regulations, PCC is implicitly included. For our purposes when specifying PCC requirements, a useful definition can be found in a new standard, EN 17398:2020, *Patient involvement in health care – Minimum requirements for person-centered care*, where a shift is attempted, away from a conventional medical approach in which the patient is a passive target of medical and/or care intervention, to an explicitly PCC approach in which:

- every patient’s resources, interests, needs and responsibilities are acknowledged and endorsed in situations of concern to them, and
- the patient takes an active part in their care, decision-making processes and self-care.

PCC emphasizes co-production of care. Particularly in times of stress such as the current COVID pandemic, it is important not to set more conventional medical health care and PCC in opposition to one another – "competing" for limited resources – but rather see them as complementary and of mutual benefit [11, 45]. For instance, an enduring personal relation between patient and health staff will provide support not only at a critical moment of intervention, but also before and after, whatever the outcome for that patient. A key aspect is the relative importance a patient attaches to various outcomes and processes having a large, if not determining, influence on decision-making in patient-centered care. Berwick [11] and Kaltoft et al. [45] also emphasize that the relevant preferences are those of the individual patient facing a decision, as opposed to the average preferences of a group of patients with the same condition or the preferences of the health professional involved in the decision.

Obviously, the umbrella term "patient centered care including physical, psychological and social integrity" (this phrase appears in EN15224 [25] which uses the term "*patient* centered care" without further definition) needs to be broken down into more specific quality characteristics (task difficulty, person ability and the like) for PCC, as will be exemplified in the later section 11.1.5.2 *PCOs: Example neuropsychological cases*. A further key aspect of PCC, not clearly covered in EN15224 but in EN 17398, is the partnership guided by the patient's narrative and characterized by dignity, compassion and respect between the patient and health care professionals [24], which can also be resolved into a number of specific quality characteristics (quality of care, person leniency), as exemplified in the later section PCOs: Example patient participation.

Apart from and together with legislation and standards, expert and end-user opinion and advice are of course valuable when defining health care. For example, the OMERACT collaboration has formulated concepts, core and domains for outcome measurement in health intervention studies [14]. A person-centered approach has even recently entered the long-established field of laboratory medicine, where one has highlighted:

> Effective collaboration with clinicians, and a determination to accept patient outcome and patient experience as the primary measure of laboratory effectiveness

in the recommendations of the IFCC Task Force on the Impact of Laboratory Medicine on Clinical Management and Outcomes (TF-ICO) [29].

11.1.3 A: Entity Attribute Description and Specification

In what follows, we will examine in turn each of the three main stages in the measurement process (Fig. 11.1a,b below) – from (*A*) an object's entity, through (*B*) measurement with an instrument to (*C*) response as registered by an operator.

The various quality characteristics of the entity, *A*, (attributes such as the quality characteristics for care services mentioned in legislation, various standards and in expert and end-user advice, first section), need to be identified, described,

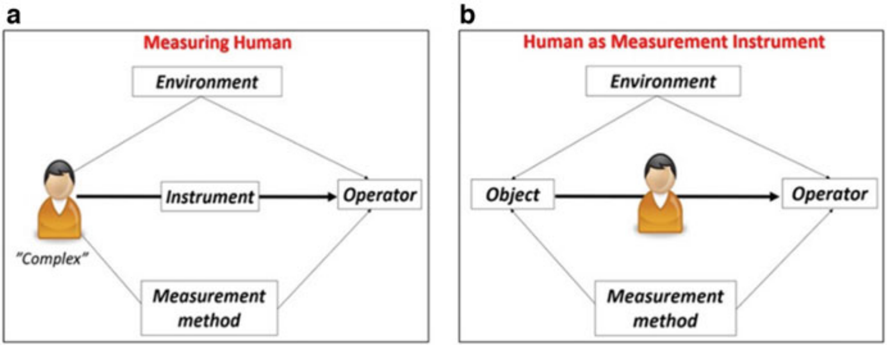


Fig. 11.1 (**a**, **b**) Measuring humans in MSA. (Adapted from [69, 72])

measurable, predictable and prioritized when formulating the overall entitic construct ("dedicated quantity" [22] or "quantity of an entity" [27]) for quality assurance. While the first section also gave some examples of typical constructs for PCC, this section indicates a number of tools and methodologies to be used when making valid construct identification.

A 'quality characteristic' attributed to any entity intended to be assessed may be – as in statistics – either a measure of:

> 'location', e.g., a direct quantity of the entity, such as the mass of a single object or ability of an individual patient, an error in mass or ability (deviation from a nominal value), or an average mass or ability of a batch of objects or cohort of patients
>
> 'dispersion', e.g., the standard deviation in mass or ability amongst a batch of objects or cohort of patients

The all-important definition of what is actually intended to be quality-assured will be in most cases a defining moment, literally speaking. Among the requirements of validity and reliability in both the physical and social sciences [84], of particular interest at this stage are: *content validity*: degree to which all necessary items are included in test; and *construct validity*: convergency and discrimination to related and unrelated measures, respectively.

In most cases PCOs can be resolved into pairs of attributes, sometimes referred to as coupling attributes or item: probe pairs: (i) a care service attribute – such as the difficulty of a task or quality of providing care – and, respectively, (ii) an attribute of the person receiving care – such as the ability of each care recipient or the leniency (how easily satisfied a person is). As will be described more in the section 11.1.5.1 *PCC including physical, psychological and social integrity*, task difficulty or quality of care, δ, and person ability or person leniency, θ, is determined together in a psychometric logistic regression to cohort performance response data obtained typically with tests of performance, questionnaires and rating scales where the measurement object is an "item" and the measurement instrument (typically here a person) is the "probe" or "agent" in a measurement system model to be described in Fig. 11.1a, b.

At this first step, there should be an ambition to capture unconditionally as many relevant and valid product and process aspects as possible – no amount of sophisticated analyses subsequently will be able to compensate for a component missed here. When identifying the processes essential for quality assurance in production, design of experiments is a tool of traditional statistics that has been employed for decades [65]. Design of experiments 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 not only in manufacturing but also more widely throughout the physical and social sciences.

In line with PCC, i.e., the aim to include the patient's narrative, and in order to develop meaningful PCOs for the patient, an example of a useful method for capturing a broad spectrum of experiences of the target group is the critical incidents' technique (CIT). A critical incident is an event that is particularly satisfying or dissatisfying, as typically identified using content analysis of stories or vignettes [12, 28], rather than quantitative methods in the data analysis.

Structuring, prioritizing and choosing among the many entity characteristics the end-user might mention may be done in different ways. In an ideal world, when formulating the construct it should relate to an ordinal theory, i.e., a prior understanding of what it means for the target group to progress from low to high ability and items correspondingly represent a harmonized hierarchy for task difficulty [17]. Some criteria for prioritizing amongst identified entity characteristics are needed when bringing structure, be it 'tailor-making' manufacturing or 'production' anywhere in the social and physical sciences [75]. An example of this in PCC is to ask patients to rate the relative importance of each task as an aid to ranking and prioritization. Tools for this include: Importance-Performance Analysis [19], where a simple experiment might be to ask a person to place task or situation cards on a table, where the position of each card is determined on the vertical axis in terms of perceived difficulty of performing that specific task and location of the card on the horizontal axis indicates how important the performance of the task is rated. The results of this simple investigation can be analyzed further using logistic regression.

The Activity Inventory tool [54] is another approach to structuring constructs, where a bank with items describing everyday activities is arranged within a hierarchical framework. At the top level of the hierarchy, activities are organized by Massof et al. [54] according to the objective they serve (Daily Living, Social Interactions, or Recreation). A selection among the (several hundred) identified items is made; a few items are classified as "Goals", which describe what the person is intending to accomplish (e.g., prepare daily meals, manage personal finances, entertain guests). The remaining items in the bank, grouped under the Goals, are classified as "Tasks". Tasks in the vision-related studies of Massof et al. [54] describe specific cognitive and motor activities that must be performed to accomplish their parent Goal (e.g., cut food, read recipes, measure ingredients, read bills, write checks, sign name). Massof and Bradley [55] report the estimation of the utility (i.e., a variable representing value, benefit, satisfaction, happiness, etc.) assigned by each patient to a hypothetical risk-free intervention that would facilitate each of the identified important and difficult Goals in the Activity Inventory for that patient.

Whatever technique is used to capture every essential aspect to be included in the construct of interest will enable a valid and reliable formulation of a construct specification equation (CSE) to be used throughout a measurement task, as described further in the section 11.3.1 *Metrological references for comparability via traceability* and by Melin & Pendrill [61].

11.1.4 Quality Assurance in Healthcare Service Provision

Following definition of the entity and its quality characteristics to be assessed for conformity with specified requirements, the quality loop progresses, via stages such as product design, procurement, through actual production to delivery of care and final use. The healthcare service quality assurance norm EN 15224 [25] stipulates – in 7.5.1 *Control of production and service provision* – that:

> The organization shall plan and carry out production and service provision under controlled conditions.

Such conditions shall include availability of information describing health care service characteristics, of necessary work instructions and of monitoring and measuring equipment as well as implementation of monitoring and measurement and of health care service release, delivery and post-delivery activities.

The specific aim of keeping product on target in conformity assessment involves setting specification limits on the quality characteristics. Here we also see how, as in all ISO 9000-like standards, measurement is included explicitly as part of the "production" process, as will be dealt with in more detail in the later sections of this chapter. Typical decision rules in conformity assessment are of the form: "Conformity to requirements is assured if, and only if, the uncertainty interval of the measurement result is inside the region of permissible values R_{PV}" (ISO 10576-1 [37]), where that region is bounded above and below specification limits, by regions of non-permissible values, R_{NV}. Note that entity specifications are normally set on the basis not only of what can be practically and economically produced, but also ultimately on what the consumer or other end-user requires in terms of product characteristics, as described above in the section 11.1.2 *Quality assurance in person-centered care. Design of experiments*. We return to decisions of conformity at the end of this chapter when we close the quality loop.

The quality characteristics and their entities (the health care services) have a different character to those associated with the more technical 'production' of care – the EN15224 standard [25] mentions several: the handling of material products such as tissue, blood products, pharmaceuticals, cell culture products and medical devices – which are not in focus in the health care service standard as they are regulated elsewhere. Alongside traditional quality-assurance of products and processes of this more "objective and quantitative" kind, many measurements in healthcare (e.g., of care services) of interest here are made with questionnaires and categorical observations.

11.1.5 Examples of Quality Characteristics and Specification Limits for Person-Centered Care

PCC, with its emphasis on the patient's unique and holistic properties brought into a partnership with health care professionals, contrasts with more traditional disease control programs. In practice, this has the consequence that the quality characteristics of typical PCC attributes are often less quantitative and more subjective – typically the response of a patient to a task – than more familiar technocratic indicators of medical signs and healthcare production measures. This can make the identification of quality characteristics for PCC, such as "equity" and "patient involvement", a challenge to health care professionals. Despite this, it may well turn out that PCC in the long run will be the most effective in providing health proactively than merely retroactively "fighting the fire".

11.1.5.1 PCC Including Physical, Psychological and Social Integrity. Specification Limits, Counted Fractions, Ability, Difficulty and the Psychometric Rasch Measurement Theory (RMT) Approach

Despite their apparent subjective character, PCC attributes can in many cases nevertheless be quantified and assessed objectively for conformity. Examples of specification limits on PCOs could be: the patient should score at least 60% on a memory test (section 11.1.5.2 *PCOs: Example neuropsychological cases*) or a patient participation survey (section 11.1.5.3 *PCOs: Example patient participation*). A couple of tools are needed to handle such data in PCC:

Firstly, it is known since the time of Pearson (1897, cited in [2], and in [91]) that raw data such as the response score, $P_{success}$, of a person to a test (perhaps administered with a survey questionnaire with a finite number of response categories) lie on an ordinal scale (what is called 'counted-fraction', bounded by 0% and 100%) which is not linear and where most of the regular tools of statistics are not directly applicable. An example will be given in Fig. 11.4 which shows the increasing non-linearity of raw data at scale extremes which, if not recognized and corrected for, will lead to incorrect decisions about healthcare. It is like making a measurement with a bent ruler ([75], pp. 88 ff). As recommended long ago by Aitchison in the 1980s and even earlier by Rasch [83], such counted-fraction data need first to be transformed onto a linear, quantitative scale – using log-ratios for instance – before attempting any statistical analysis. Thereafter one can transform back to the observation space. It is straightforward to convert a raw-score specification $P_{success}$= 60%

limit to a corresponding log-ratio specification limit $z = 0{,}41$ logits using the equation: $z = \ln(\frac{P_{success}}{1-P_{success}})$. The centre of the logistic scale, $z = 0$, is clearly at the raw data point $P_{success} = 1 - P_{success} = 50\ \%$.

Secondly, it is well-known that the success rate $P_{success}$ in the response to a test or classification depends not only on the test person's ability θ but also on how difficult, δ the test is. The same probability of success can be obtained with an able person performing a difficult task as with a less able person tackling an easier task. Any measurement is indirect, that is, there is always the need to have the intervention of a measurement system in order to observe any object of interest, as modelled with MSA.

To the extent that care, including PCC, makes both diagnoses and attempts to cure illness, then relevant quality characteristics are, respectively, the actual level of patient ability θ at any one diagnosis, as well as the change in patient ability following an intervention, $\Delta\theta$. Further discussion can be found for example in the standard ISO 18104 [40] (and in the section 11.4.2.1 *Rater constructs*), which aims at providing a comprehensive categorical structure or domain concept model for nursing diagnoses and nursing interventions.

The same approach can be applied when considering responses to a care participation test, as will be described further below, where the corresponding care service and care recipient attributes are, respectively, care quality and recipient leniency (how easily satisfied a person is).

How to formulate an MSA for categorical observations needs to be a key focus in PCC. Although Rasch [83] did not use those terms, an early form of his model, in response to contemporary demands for individual measures (section 11.2.3 *A way forward for measuring PCOs: Human as a B: Measurement Instrument*), is very much an MSA formulation [75] which posits that the odds ratio of successfully performing a task is equal to the ratio of an ability, *h*, to a difficulty, *k*:

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

(Rasch [83] used the person attribute "inability" instead, given by h^{-1}, which can be written:

$$\theta - \delta = \log\left(\frac{P_{success}}{1 - P_{success}}\right) \qquad (11.1)$$

where the test person ("agent") ability, $\theta = \ log\ (h)$, and task ("object") difficulty, $\delta = \ log\ (k)$ (or other item:probe pairs), can be evaluated by logistic regression to the score data in terms of the probabilities $q\ (P_{success})$ which quantify the response of the measurement system to a given stimulus (object difficulty). The original Rasch [83] formulation of Eq. 11.1 refers to a Poisson probability distribution $p(x) = \frac{e^{-\lambda}\cdot\lambda^{x}}{x!}; x = 0, 1, \ldots,$ 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 [65]. The parameter λ is equal directly both to the mean and variance of the Poisson distribution and in Rasch's [83] model $\lambda = h^{-1} \cdot k$. In accord with MSA, the object of measurement is, simultaneously, the object itself as well as an essential element of the measurement process. So, while here the discussion concerns mainly the object of interest (e.g., PCOs as in the next sections), the rest of this chapter will deal more with the quantities as measured. Both the mean and variance of this Rasch distribution will be considered when dealing, respectively, with the measured person abilities and task difficulties on the one hand, and the uncertainties in each of these on the other. RMT, as will be seen, is not only a statistical approach but is, importantly, a metrological approach to person-centered measurement ([69, 72] and Fig. 11.1a, b).

11.1.5.2 PCOs: Example Neuropsychological Cases

A quality characteristic in a neuropsychological memory test regularly used in clinics monitoring neurodegeneration is the ability, θ of a patient to remember a particular sequence, for instance of blocks tapped, digits or words in a list. In the latter section of this chapter we will give examples of actual data and analyses, with more detailed accounts given in the accompanying chapter by Melin & Pendrill [61].

A first question is: Is patient ability θ a quality characteristic of person-centered healthcare? We would argue that the characteristic is certainly person-centered, perhaps more so than characteristics – such as brain volume, protein concentration, etc. – of the traditional technocratic biomarker approach, where an impressive array of sophisticated instruments and theories of brain-functioning has had decades of research dedicated to it. At least a patient's performance is directly relevant and certainly not a "surrogate" for biomarker levels. Secondly, since the aim of healthcare is to maintain and preferably improve health, then any change in patient ability $\Delta\theta$ is a pertinent quality characteristic. Our view can be compared with that of Walton et al. [94] who write:

> clinical outcome assessments (COAs) . . . include any assessment that may be influenced by human choices, judgment, or motivation. COAs must be well-defined and possess adequate measurement properties to demonstrate (directly or indirectly) the benefits of a treatment. In contrast, a biomarker assessment is one that is subject to little, if any, patient motivational or rater judgmental influence.

11.1.5.3 PCOs: Example Patient Participation

A couple of years ago, in a Health Foundation report, [20] pointed out that there is neither a universally agreed definition of PCC, nor a golden standard that can be used for measuring all aspects of PCC. Likewise, earlier critique emphasized that few of existing scales were based on a solid PCC theoretical framework [23]. There is, however, a lot of research and proposals ongoing about how to define and measure experiences of PCC, e.g., to assess a subdomain of PCC such as patient participation.

In an ideal world, the ordinality of quantity scales (as described above in the section 11.1.5.1 *PCC including physical, psychological and social integrity*) should be recognized at an early phase of construct specification. This was, however, not the case with the *Patient Participation in Rehabilitation Questionnaire* (PPRQ) [51]. A subsequent analysis according to RMT [62] has identified a reasonable item hierarchy of what it means to go from lower quality characteristics, where the lower levels include being respected as a person and being given information, while the middle levels include shared decision making and care planning and the higher include being empowered and motivated.

There are striking similarities between items and the hierarchical levels identified about the same time in two other independent developed questionnaires, *Patient Preference for Patient Participation tool* (The 4Ps) [52] and the *Person-Centered Care in outpatient care in rheumatology* (PCCoc/rheum) [8]. Consequently, it appears promising in the understanding of the construct PCC/patient participation from a patient perspective on these scales.

11.2 Benefits of Combining Rasch Measurement Theory (RMT) and Quality Assurance

After having decided on which clinical processes are to be quality-assured in person-centered healthcare and what the principal quality characteristics associated with the processes are, the next essential steps in our quality loop are monitoring and measurement of these characteristics during the actual provision of healthcare services.

As we have seen above, clear demands about measurement can be found in quality assurance and conformity assessment, such as in legislation: ‘the quality of the healthcare shall be systematically and continuously developed and assured’ [31]. The healthcare service quality assurance norm EN 15224 [25], although perhaps not explicitly covering PCC in detail, nevertheless usefully stipulates (as other ISO 9000 norms) metrological requirements in general (section 11.2.2.1 *Requirements for traceable measurement,* below).

When making measurements at this “production” stage in the quality loop, it is important, and one of the most challenging tasks, to distinguish between measurement dispersion (due to limited measurement quality) and actual product (health care service in the present case) variation. The two types of scatter appear together in the displayed results of measurement and are easily confounded, even conceptually [75]. For instance, as a patient progresses in her health condition or as a result of care interventions, of main concern in care provision will be variations, such as changes in ability, actually caused by illness or the intervention. It is obviously essential and beneficial in providing good care to distinguish these actual changes from apparent changes registered with a less-than-perfect measurement system. Indeed, most measurements are made through the intermediary of a measurement instrument, so some account is necessary to compensate for possible imperfections and distortions in the measurement process if one wants a true picture of the object of interest – in the present case, the clinical processes. This in turn is key if measurements are ultimately to provide a reliable and valid basis for clinical decisions (see further in the last section 11.4 *Decision risks and uncertainty*).

A unique aspect when assuring measurement quality in PCC turns out to be that the many measurements made, for instance with questionnaires, tests and surveys – often made on categorical scales – are best analyzed with a measurement model where the human respondent is herself acting as the measurement instrument, as a metrological interpretation of RMT [69, 71–73, 75]. An MSA with a human measurement instrument is as "person-centered" as PCC, where one is interested not only in measuring the status of a person, but also understanding what and how the person perceives and experiences. At the same time, this approach enables the benefits of drawing analogies with the well-established MSA framework from measurement and quality engineering [65] to be fully exploited, where the instrument is at the heart of any measurement system.

11.2.1 Measurement Quality-Assurance Loop

Confidence in the measurements performed in the conformity assessment of any entity (product, process, system, person, body or phenomenon) can be considered sufficiently important – as in the present case of care – that the measurements themselves will be subject to a "product" quality loop as a kind of metrological conformity assessment 'embedded' in any product Conformity Assessment.

A measurement quality-assurance loop is formed by the steps a. . .f (also illustrated in figure 2.1 of [75]) is of course analogous to a product quality-assurance loop, where the "product" is a measurement result:

(a) Define the entity and its quality characteristics to be assessed for conformity with specified requirements (section 11.1.3 *A: Entity attribute description and specification*).
(b) Set corresponding specifications on the measurement methods and their quality characteristics (such as maximum permissible uncertainty and minimum measurement capability) required by the entity assessment at hand (section 11.2.2 *Measurement specifications*)
(c) Produce test results by performing measurements of the quality characteristics together with expressions of measurement uncertainty (section 11.3 *Metrological references for comparability via traceability and reliable estimates of uncertainty*).
(d) Decide if test results indicate that the entity, as well as the measurements themselves, is within specified requirements or not (section 11.4 *Decision risks and uncertainty*).
(e) Assess risks of incorrect decisions of conformity.
(f) Final assessment of conformity of the entity to specified requirements in terms of impact.

11.2.2 Measurement Specifications

In order to assess an entity according to product specifications, it will often be necessary to set corresponding measurement specifications (step b, above). Much can be gained if one can plan measurements proactively for “fitness for purpose”, encompassing considerations – prior to measurement – of aspects such as the measurement capability needed when about to test a product or process of a certain production capability, in the same way as design of experiments was used to find the principal input factors determining the output of “production” [73].

Models of measurement will need to be able to deal with:

- The actual result of measurement, that is, estimating the measurand or quality characteristic of the entity of interest
- Propagation of measurement bias and other errors through the measurement system
- Propagation of variances through the measurement system

The two principal hallmarks of metrology (quality-assured measurement) – namely, traceability and measurement uncertainty, respectively – are the two aspects which ensure that the product (in the broadest sense) itself is quality-assured, so that it will be comparable and decision risks will be quantified. Assuring the quality of measurement in terms of the respective measures of “location” and “dispersion” will in turn support corresponding quality assurance of product – in the present case the requirements on the attributes of PCC.

As in all measurement, there are requirements when assuring quality on both metrological traceability and general risk management. These two metrological aspects are of course key to ensuring quality in PCC for several of the PCC quality characteristics listed in the section above 11.1.2 *Quality assurance in person-centered care. Design of experiments*, such as “equity” (i.e., a patient can expect the same quality of case wherever it’s provided) and “patient safety” (i.e., risks of incorrect decisions about care are proactively assessed). When assuring quality in PCC, there are some special challenges to be met when fulfilling both these sets of requirements.

11.2.2.1 Requirements for Traceable Measurement

The healthcare service quality assurance standard EN 15224 [25] (as in all ISO 9000 like norms) stipulates – in 7.6 *Control of monitoring and measuring equipment* – that:

> The organization shall establish processes to ensure that monitoring and measurement can be carried out and are carried out in a manner that is consistent with the monitoring and measurement requirements.
>
> Where necessary to ensure valid results, measuring equipment shall:

> a) be calibrated or verified, or both, at specified intervals, or prior to use, against measurement standards traceable to international or national measurement standards; where no such standards exist, the basis used for calibration or verification shall be recorded . . .
>
> NOTE 2 For further information see EN ISO 10012:2003 Measurement management systems – Requirements for measurement processes and measuring equipment [36].

It can be noted that, to date, there are few established measurement standards for attributes typical of PCC.

11.2.3 A Way Forward for Measuring PCOs: Human as a B: Measurement Instrument

Metrological traceability and uncertainty are equally pertinent to assuring measurement quality in PCC, as they are to more traditional quality assurance contexts. While superficially similar, there are however a number of caveats and challenges to be overcome in person-centered care. An early mention of the topic was given by Maxwell [57]:

> We may observe the conduct of individual men and compare it with that conduct which their previous character and their present circumstances, according to the best existing theory, would lead us to expect. Those who practice this method endeavour to improve their knowledge of the elements of human nature, in much the same way as an astronomer corrects the elements of a planet by comparing its actual position with that deduced from the received elements.

Rasch [83], in motivating what was to become RMT, quotes [100]:

> Recourse must be had to individual statistics, treating each patient as a separate universe. Unfortunately, present day statistical methods are entirely group-centered, so there is a real need for developing individual-centered statistics.

RMT is, importantly, a *metrological* approach to person-centered measurement [69, 72]. This contrasts with current thinking in some quarters (stemming from attitudes prior to Rasch's approach of the 1960s) that 'secondary to the scientific task is the instrumental task of quantification' [64]. Even today, accounts [53] seem still to follow this thinking, and essentially omit, to their detriment, description of a measurement system in any detail. Drawing simple analogies between "instruments" in the social sciences – questionnaires, ability tests, etc. – and engineering instruments such as thermometers does not unfortunately go far enough. Not only does the counted-fraction nature of human responses need to be compensated for, but also and the task of separating the instrument factor from the sought-after object factor has to be achieved even in qualitative, categorical responses of the measurement systems. The importance and benefits of regarding the human responder as an instrument in MSA are explained at the beginning of this section.

Attempts have been made, for instance by analytical chemists [22, 67] who have proffered a suggestion that 'examination' could replace 'measurement' in the case of nominal properties (a specific kind of categorical observation). In the example 1 of VIN §3.9 *Examination uncertainty* given by Nordin et al. [67]:

> The reference nominal property value is "B". The nominal property value set … of all possible nominal property values is {A, B}. For one of 10 examinations … the examined value differs from "B". The examination uncertainty is therefore 0.1 (10 %).

Texts such as this raise a couple of questions:

Firstly, the term "examination" in English usually means either a "detailed study" or a "test". For categorical observations, we prefer the term "classification" which is commonly used (for example in taxonomy) and can cover less detailed studies (which are nevertheless important to make) as well as not obliging us to make tests.

Secondly, performance metrics such as $P_{success}$, and misclassification probabilities, α or β – considered by Nordin et al. [67], Bashkansky et al. [9] and others as accuracy measures, often belong to the 'counted fraction' kind of data (section 11.1.5.1 *PCC including physical, psychological and social integrity*) and in general are of ordinal, rather than fully quantitative nature, and are not directly amenable to regular statistics. An uncertainty of "10%" has little meaning. There is also, again, the requirement that the distinct factors of task difficulty and person ability need to be determined separately from the aggregate raw score $P_{success}$.

Not accounting for the counted-fraction nature of human responses can have serious consequences in healthcare. An example is given by Pendrill [76] who demonstrated significant errors in previous studies of the correlation of the cognitive ability of Alzheimer disease sufferers with corresponding biomarkers when not compensating for known counted-fraction scale distortion. Another example is by Kersten et al. [46] which showed that raw data are invalid for decisions of *Minimum Clinically Important Differences* as these either under- or overestimate true changes.

In summary, the status quo concerning a metrological vocabulary which would adequately cover ordinal and nominal properties satisfactorily would need three additions to that available internationally today:

1. clear definitions of ordinal and nominal properties (which in most cases are *not* quantities: A "quantity" is a measurable property)
2. a new chapter in the vocabulary defining classification systems, analogous to measurement systems, but where the response is the probability of a "correct" assignment of an observation to a class or category
3. clearer definitions of what a measurement system is, including the quantities associated with each element of the system, and the process of restitution [85]

With the explicit aim of developing a common language and complementary methodologies for cooperation about measurement and decision-making between sociologists, physicists and others, we have recently proposed [69, 71, 72, 75] an approach that seems to be equally applicable to both physical and social measurements and combines both aspects. While it is difficult to find a single definition of “measurement” which would apply to all scales – from nominal through to ratio – [82], it does appear to be feasible to unite about a definition of “classification”, including qualitative observations typical of PCC as well as vacillation when making decisions based on quantitative measurement results in the presence of measurement uncertainty (“unknown measurement errors”).

Almost all measurements are not direct, but in most cases human beings need the help of instruments to translate measurement information from a measurement object into a human-understandable signal. Thus, a measurement system (depicted in Figs. 11.1a, b) is necessary but at the same time because it is not perfect, brings with it more or less measurement errors (noise, distortion) that must be corrected for [10, 75].

In Fig. 11.1 (a) a conventional measurement system ([1, 7], MSA *Measurement System Analysis*), perhaps used in traditional healthcare for measurement of a person’s mass, length or temperature – challenging owing to the complexity of a human being – is contrasted with Fig. 11.1 (b) of a “person-centered” measurement system suitable for PCOs, where a human being acts as a measurement instrument at the heart of a measurement system [71, 72].

Our approach [71, 72, 74, 75] identifies the test person (e.g., patient in the present context) as the instrument at the center of the measurement system (Fig. 11.1b), in accord with Rasch’s [83] intentions to ‘treat each patient as a separate universe’ and that a PCC needs person-centered metrology. Note however that, as reported already in the EU MINET project [69, 70]:

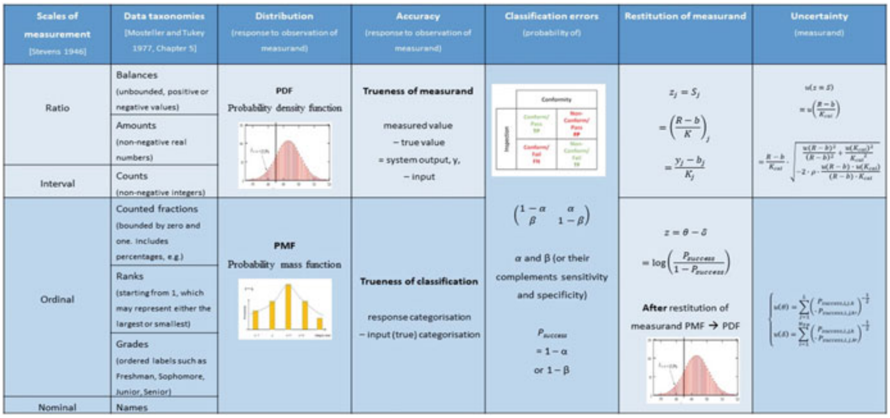
> Care has to be exercised in such studies and an overall aim is to *bridge the gap* between: Engineering tradition criticized for a far too instrumental view of operators. Humanistic and behavioral science tradition all too preoccupied with issues centered on human operators.

The MSA approach is applicable to, in principle, all scales of measurement – as *measurement* systems for the more quantitative ratio and interval scales, and as *classification* systems for the less quantitative ordinal or nominal scales, as summarized in Fig. 11.2.

11.2.4 Benefits of Analysing Response Data with RMT

Classification, where responses are put into a number of categories (class labels) – and visualized typically with a PMF (probability mass function, histogram of occupancy in a finite number of categories) – is analogous to measurement, where responses are put on a continuous scale of the measurand (quantity intended to be measured) and visualized typically with a PDF (probability density function).

Parameters used to characterize measurement systems (in MSA, including object, instrument, method, environment and operator) have often analogous meanings when characterizing classification system. For example, a “classification error” is analogous to a “measurement error” (see Fig. 11.2). Similarly, concepts such as



Scales of measurement [Stevens 1946]	Data taxonomies [Mosteller and Tukey 1977, Chapter 5]	Distribution (response to observation of measurand)	Accuracy (response to observation of measurand)	Classification errors (probability of)	Restitution of measurand	Uncertainty (measurand)
Ratio	Balances (unbounded, positive or negative values)	PDF Probability density function	**Trueness of measurand** measured value − true value = system output, y, − input	Conformity Inspection Conform/ Pass TP Non-Conform/ Pass FP Conform/ Fail FN Non-Conform/ Fail TN	$z_j = S_j$ $= \left(\frac{R-b}{K}\right)_j$ $= \frac{y_j - b_j}{K_j}$	$u(z = S)$ $= u\left(\frac{R-b}{K_{cal}}\right)$ $= \frac{R-b}{K_{cal}} \cdot \sqrt{\frac{u(R-b)^2}{(R-b)^2} + \frac{u(K_{cal})^2}{K_{cal}^2} - 2 \cdot \rho \cdot \frac{u(R-b) \cdot u(K_{cal})}{(R-b) \cdot K_{cal}}}$
	Amounts (non-negative real numbers)					
Interval	Counts (non-negative integers)					
Ordinal	Counted fractions (bounded by zero and one. Includes percentages, e.g.)	PMF Probability mass function	**Trueness of classification** response categorisation − input (true) categorisation	$\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$ α and β (or their complements sensitivity and specificity) $P_{success}$ $= 1 - \alpha$ or 1 − β	$z = \theta - \delta$ $= \log\left(\frac{P_{success}}{1 - P_{success}}\right)$ **After** restitution of measurand PMF → PDF	$\begin{cases} u(\theta) = \sum_{j=1}^{L}\left(P_{success,i,j,k} \cdot P_{success,i,j,k'}\right)^{-\frac{1}{2}} \\ u(\delta) = \sum_{i=1}^{N_{TP}}\left(P_{success,i,j,k} \cdot P_{success,i,j,k'}\right)^{-\frac{1}{2}} \end{cases}$
	Ranks (starting from 1, which may represent either the largest or smallest)					
	Grades (ordered labels such as Freshman, Sophomore, Junior, Senior)					
Nominal	Names					

Fig. 11.2 Descriptions of measurement scales and various metrological parameters (S = stimulus; R = response of measurement system; b = bias in response; K = sensitivity; y = indication). Light blue cells: interval or ratio quantities; dark blue cells: nominal or ordinal properties (not quantities)

"sensitivity", "specificity", "selectivity", "uncertainty", etc. are analogous between classification systems and measurement systems. However, as illustrated in Fig. 11.1b, the "instrument" in a classification system can be a classifier such as a person.

These analogies reflect the fact that the more quantitative properties (ratio, interval) include at the same time the more basal characteristics of the less quantitative properties (ordinal, nominal). But the usual tools of statistics **cannot** be directly used for classification system response since distances on ordinal and nominal scales are not known. Ratio and interval scales belong to quantities (which can be measured) while ordinal and nominal properties are **not** quantities, but are based on categorical observations (i.e., classifications). However, as will be described below and in Melin & Pendrill [61], through measurand restitution, ordinal and nominal properties can be transformed into quantities.

For most people, a measurement does not stop with a "measurement result", that is, a value of the measurand (after measurand restitution), but measurement is done for some, third party reason, such as conformity assessment of products (entities).

In fact, continuing a measurement to include a decision – such as the entity is approved or rejected with respect to a specification limit – turns the measurement system into a classification system! Thus, risks of decision-errors (vacillation – arguably closer to everyday "uncertainty") arising from finite measurement uncertainty will belong to ordinal properties.

In this case, the MSA is very useful to describe classification system performance in terms of the ability of a classifier (instrument) to make a rating; the level of difficulty (task object); the ability of the operator rate the rater. This is discussed further below.

It is fitting that MSA is the chosen technique to interpret RMT [83] metrologically, developed as it has been for quality-assurance of measurements in the "field" or the "workshop floor" [1, 7]. In contrast to the well-controlled conditions of the measurement laboratory, the majority of instruments (both traditional, engineering kinds as well as the PCOs, humankind considered here in PCC) are used to make measurements where the surroundings may have considerable influence on every element of the measurement system. A patient undergoing a psychometric test, for instance, might easily be perturbed in her response by some disturbance from the environment in the "field" of the clinic, as modelled with MSA.

Other factors determining measurement outcome to consider are (i) possible interactions between the instrument (person) and the object (task performed) – similar to instrumental 'loading' in engineering measurement systems ([10, 71–73], Chapter 2 in [75]) and (ii) the particular measurement method chosen. All these factors are included in our MSA approach (despite statements to the contrary in the recent literature [93]).

The measurement system approach (Fig. 11.1) is essential in all sciences in obtaining both aspects of metrology – traceability and uncertainty:

Metrological standards, reflecting fundamental symmetries which provide minimum entropy and constants for measurement units (Chapter 3 in [75, 77]), traceability to which can enable the comparability of measurements (and by extension even the intercomparability of products and services of all kinds), can only be established if one can separate out – with a process called measurand restitution – the limiting factors of the measurement system used to make measurements from the response of the system to a given stimulus from the measurement object (section 11.3 *Metrological references for comparability via traceability and reliable estimates of uncertainty*).

Less than complete information about a system leads to uncertainties, leading to incorrect decision making, for example approval of an incorrect product. Formulation of a performance measure, i.e., how well a measurement system performs an assessment – actually measurement uncertainty – seems to be treated with similar (categorical) methods, whether it is "instruments" – questionnaires, examinations etc. – in social science or in assessing how well a measuring instrument shows if an item or product is within or outside a specification limit (section 11.4 *Decision risks and uncertainty*).

Using MSA to interpret metrologically the RMT [83] adds extra insight into the meaning and impact of the various terms in Eq. (11.1). For instance, while mathematically the pair of attributes (δ, θ) appear symmetrically (apart from a difference in sign) in the logistic function, metrologically the role played by the object attribute and the instrument attribute, respectively, are very different, as will be illustrated in the rest of this chapter.

11.3 Metrological References for Comparability via Traceability and Reliable Estimates of Uncertainty

Much of the established metrological terminology of physical measurement carries well over into measurement in the human sciences (chapter 4 of [75], and Fig. 11.2), such as PCOs. There are however caveats, as mentioned (e.g., in the above section 11.1.5.1 *PCC including physical, psychological and social integrity*):

Firstly, that data obtained from the response of a measurement system where a human is the instrument are often not themselves directly amenable to the usual statistical tools (e.g., calculating a mean or standard deviation) owing to the ordinal or nominal character of the raw data. Rather, the raw data are classifications.

Secondly, and often forgotten, as with any data obtained via an instrument, metrological traceability and reliable assessment of uncertainties and decision-risks with human science data require a proper measurement system analysis, where there are preferably clear and separate estimates of the contributions of each element of the measurement system – instrument, operator, environment and measurement method (Fig. 11.1) – to the overall response when measuring a certain object.

The recommendation to deal with these caveats is that, instead of attempting to treat raw data with invalid statistical tools, one transforms the classification system response $P_{success}$ (e.g., the probability of making a correct decision or performing a task of a certain difficulty) lying on a less quantitative ordinal or nominal scale, onto a more quantitative interval or ratio scale by applying the RMT formula (Eq. 11.1). (Chapter 5 of [75] presents appropriate tests of the validity of the transformation.)

11.3.1 Metrological References for Comparability via Traceability

The use of a measuring stick, such as a ruler marked with a scale for length, is familiar to most when determining how many times a unit fits into the quantity to be measured. Maxwell wrote [58]:

> Preliminary on the Measurement of Quantities.
>
> 1) EVERY expression of a Quantity consists of two factors or components. One of these is the name of a certain known quantity of the same kind as the quantity to be expressed, which is taken as a standard of reference. The other component is the number of times the standard is to be taken in order to make up the required quantity. The standard quantity is technically called the Unit, and the number is called the Numerical Value of the quantity. There must be as many different units as there are different kinds of quantities to be measured. . . .

In its logistic regression form, the ‘straight ruler’ aspect of the RMT formula, i.e., Eq. (11.1), has been described by Linacre and Wright [50] in the following terms:

> The mathematical unit of Rasch measurement, the log-odds unit or ‘logit’, is defined prior to the experiment. All logits are the same length with respect to this change in the odds of observing the indicative event.

The RMT approach goes further in defining measurement units [33] since it uniquely yields estimates ‘not affected by the abilities or attitudes of the particular persons measured, or by the difficulties of the particular survey or test items used to measure,’ i.e., specific objectivity [34]. The RMT [83] approach is thus not simply mathematical or statistical, but instead a specifically metrological approach to human-based measurement.

Note that the same probability of success can be obtained with an able person performing a difficult task as with a less able person tackling an easier task. The separation of attributes of the measured item from those of the person measuring them – visualized with MSA (Fig. 11.1) – brings invariant measurement theory to psychometrics. Fisher’s [26] work on the metrology of instruments for physical disability was one of the first to demonstrate the concepts of intercomparability through common units, commonly referred to today as “item banking” [80]. Stone [90] for instance has written:

> Item and person separation statistics in Rasch measurement provide analytic and quality control tools by which to evaluate the successful development of a variable and by which to monitor its continuing utility. . . . The resulting map is no less a ruler than the ones constructed to measure length.

Having enabled with RMT a set of metrological references, e.g. for task difficulty, one can then proceed to set up a scale (analogous to conventional measurement etalons (Fig. 11.3)) which is delineated by measurement units where any measured quantity, $\delta_j = \{\delta_j\} \cdot [\delta]$, is the product of a number { } and a unit denoted in square brackets [ ], according to Maxwell’s [58] text quoted above.

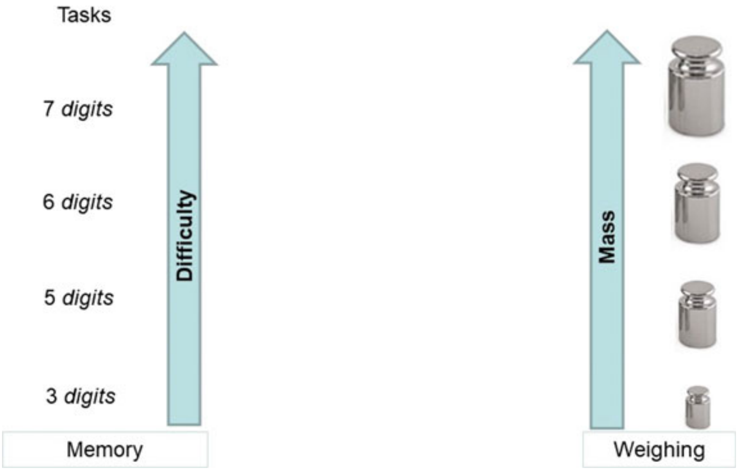


Fig. 11.3 A set of tasks of increasing difficulty (recalling a series of digits in the memory test Digit Span Test, DST) as metrological references analogous to a set of increasingly heavy mass standards

This step of establishing metrological references is enabled by combining a procedure to transform qualitative data (i.e., classifications) to a new ‘space’ (in the present case, through restitution, to the space of the measurand), together with ability of RMT to provide separate estimates of measurement and object dispersions in the results when a human acts as a measurement instrument.

This new approach to the metrological treatment of qualitative data differs from others in that the special character of the qualitative data is assigned principally not to the measurand (object entity characteristic) but to the response of the classification system. (A person-centered measurement process, where the human responder is the instrument at the heart of the measurement system, places the focus clearly on what and how the person perceives and experiences, analogous to the common expression: ‘Beauty is in the eye of the beholder’.) Using RMT in the restitution process establishes a linear, quantitative scale for the measurand (e.g., for a property such as task difficulty) where metrological quality assurance – in terms of traceability and uncertainty – can be performed.

11.3.1.1 Reference Measurement Procedures: Construct Specification Equations as Recipes for Traceability

“Recipes” to define measurement units in the social sciences, analogous to reference measurement procedures in analytical chemistry and materials science, appear promising as a viable procedure to establish metrological references in fields such as PCC where one does not enjoy access to universal units of measurement as in Physics (section 4.4.3 in [61, 75]). A condition is of course that the principal caveats of such data have been dealt with adequately, as described above.

What we earlier referred to, literally speaking, as a defining moment, was the all-important definition of which construct is actually intended to be quality-assured (section 11.1.3 *A: Entity attribute description and specification*). After identifying in that process all significant variables which can explain a particular construct, a construct specification equation (CSE) $\delta = f[x_1 \cdots x_m] = \sum_{k=1}^{m} \beta_k \cdot x_k$, can be developed. This procedure can be done for both item (object) attribute δ and person (instrument) attribute θ. A step-by-step description of ways of formulating CSEs and examples are given in the case of memory in the accompanying chapter by Melin & Pendrill [61] in this book.

11.3.2 Errors and Uncertainties in PCOs

Some typical data shown in Fig. 11.4 help describe errors and uncertainties in PCOs, where the example data set is described by Linacre (WINSTEPS ® [96]) as follows:

> 35 arthritis patients have been through rehabilitation therapy. Their admission to therapy and discharge from therapy measures are to be compared. They have been rated (1–7, although 6 or 7 are not used in the current example) on the 13 mobility items (e.g., the task of ‘eating’) of the Functional Independence Measure (FIM™):

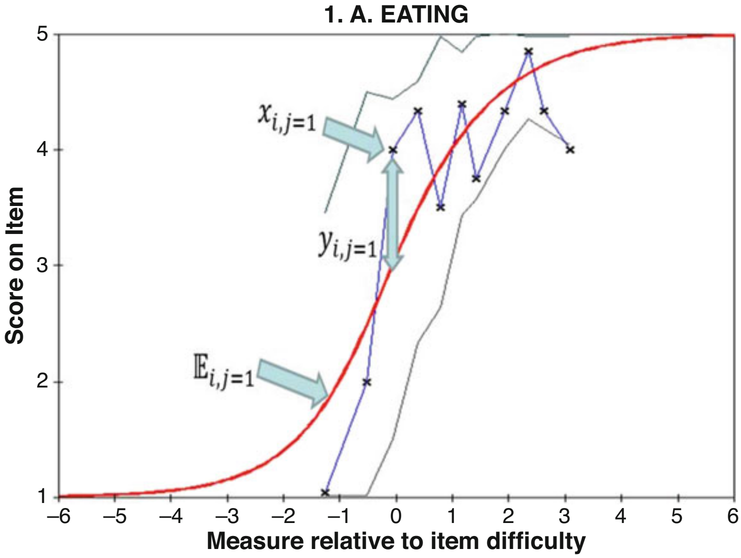


Fig. 11.4 Residuals of logistic regression as differences $y_{i,j} = x_{i,j} - \mathbb{E}_{i,j}$ between each observed score $x_{i,j}$ (i.e. classification) and the expected score $\mathbb{E}_{i,j}$ (i.e. the estimated measurand) for one item (task of 'eating') and across the cohort of test persons, *i,* in the example data set 'EXAM12' provided with the WINSTEPS® program

1. 0% Independent
2. 25% Independent
3. 50% Independent
4. 75% Independent
5. Supervision
6. Device
7. Independent'

Such a plot illustrates many of the essential aspects to consider when dealing with data based on classifications. The solid red line, with the characteristic ogive form, shows the result of a least-squares regression fit to the logistic formula Eq. 11.1 to the raw data for one particular item (task of 'eating') across the cohort of patients. How well the fit is made mainly determines the uncertainties in patient ability estimates.

Ogive curves of the kind exemplified in Fig. 11.4 show how the success rate $P_{success}$ varies from 0% to 100% across the range of abilities of individuals in the test cohort. With a simple binary – "yes-no" – score without uncertainty, the curve would instead be a sharp step function, where 0% to the left of the zero of the logistic ability scale switches immediately to 100% to the right of zero. The example of Fig. 11.4 is the case where raw data scoring is instead divided into a finite number (five in the present case) of response options (categories) for classifications, which when fitted to the logistic Eq. 11.1 leads to a smooth ogive curve instead of a step.

11.3.2.1 The Peculiar Sensitivity of a Human as a Measurement Instrument

Apart from the reservations already expressed about considering a human being as an “instrument” (section 11.2.3 *A way forward for measuring PCOs: Human as a B: Measurement Instrument*), with all the complexities that human behavior implies, there are some more specifically technical issues to deal with. (Discrimination will be tackled later.)

As with all counted-fraction data (bounded by 0% and 100%), the raw data scale (*y*-axis of Fig. 11.4) becomes strongly non-linear when approaching the lower and upper ends of the scale. This is simply stating that, at either scale extreme, the response has to be one response or the other – either “yes” or “no” – in this binary, two-category case. Don’t expect a Normal distribution of responses [95]. Attempts to use raw data in correlation studies (e.g., against biomarker concentration) or fit residual studies will certainly not be valid for test persons who find this item (task of eating) either too easy (those to the right of the plot) or too difficult (to the left of the plot) [76].

The sensitivity of the instrument (person) to changes, in contrast, varies rapidly at the “sweet spot” at mid-range where Rasch’s models for measurement are most revealing ([75], p. 178). This spot is at the zero of the horizontal x-axis of person ability in Fig. 11.4 is where the ability equals the task difficulty, that is, where there is a probability of succeeding with the task, $P_{success} = 1 - P_{success} = 50\%$. An “ordinary” measurement instrument in physics and engineering has a sensitivity (i.e., the relation between the output response to a given input) more or less constant across a range of measured values. The sensitivity of the instrument in PCOs (the person) is very different: sensitivity is greatest at mid-range, while at either end of the scale, well away from mid-range, the Rasch response model indicates that large excursions in test person ability at either extreme have negligibly small effects on the raw score.

This means, for instance, that studies of healthy cohort members (such as when researching early detection of disease degeneration) will be especially challenging. There seems to be no alternative to transforming the raw data – where the Rasch transformation ‘stretches out’ the non-linearity – if one is to have any chance of reliably resolving person ability differences. Apart from the logit approach of RMT, there are other possible transformations: Tukey [91] mentions, for example, arc-sine “anglits” and Normal ruling “normits” or “probits”. However, it seems that the peculiar sensitivity makes the choice of restitution transformation less critical.

This peculiar sensitivity will also provide a special “flavor” to how measurement uncertainties propagate through the measurement system, from object through instrument to response, as discussed next.

11.3.2.2 Errors and Uncertainty. Reliability

In general, on inspection it will be found that the estimated person attribute θ (e.g., a level of ability of a particular person or instrument) differs, because of limited measurement reliability, from the 'true' θ', with an error ε_θ:

$$\theta = \theta' + \varepsilon_\theta$$

and the limited reliability evident in estimates of task difficulty:

$$\delta = \delta' + \varepsilon_{\delta.}$$

Such deviations arise, at least in part, because the measurement instrument (the person) used to 'probe' the object or item is not perfect. While every effort should be expended by the metrologist to evaluate (e.g., through calibration), correct for and reduce these measurement errors, there will always be – because resources are limited – some measurement errors which remain unknown. Measurement uncertainty is an estimate of these unknown measurement errors.

As in all statistics, the more data points (degrees of freedom) one has, the better the reliability in cases where statistical noise is a dominant source of uncertainty. In logistic regression, as in a simple calculation of a mean value, the reliability variance increases linearly in proportion to the number of data points. This is captured in the Spearman-Brown formula (Eq. 11.2, [88, 89]), which relates reliability coefficients, R_C and R_T, for the current (C) and target (T) tests to the corresponding test lengths, L_C and L_T, that is, the number of instruments (persons) or objects (items), is:

$$L_T = L_C \cdot \frac{R_T \cdot (1 - R_C)}{R_C \cdot (1 - R_T)} \tag{11.2}$$

where a reliability coefficient (R_z) for an attribute, z, is calculated as:

$$Reliability, R_z = \frac{True\ variance}{Observed\ variance} = \frac{Var(z)}{Var(z')} = \frac{Var(z') - Var(\varepsilon_z)}{Var(z')} \tag{11.3}$$

In psychometrics a measurement reliability coefficient (calculated with Eq. (11.2)) of 0.8 – corresponding to a measurement uncertainty of about one-half of the measured dispersion – is considered acceptable for so-called high stakes testing [48]. A factor one-half is also a pragmatic limit [73] to limiting the impact of decision risks. Such reliability limits will of course constitute minimum requirements on sample size (number of test persons) and test lengths in random control trials, often formulated in regulations (e.g., by the regulators FDA and EMA).

Although often used when estimating measurement uncertainties, a Bayesian approach has been criticized by Cramér [18], who wrote:

> In a case where there are definite reasons to regard ε (i.e., measurement error) as a random variable with a known probability distribution, the application of the preceding method [Bayes theorem) is perfectly legitimate and leads to explicit probability statements about the value of ε corresponding to a given sample. However, in the majority of cases occurring in practice, these conditions will not be satisfied. As a rule, ε is simply an unknown constant and there is no evidence that the actual value of this constant has been determined by some method resembling a random experiment. Often there will even be evidence in the opposite direction, as for example in cases where the ε values of various populations are subject to systematic variation in space or time. Moreover, even when ε may be legitimately regarded as a random variable we usually lack sufficient information about its *a priori* distribution.

Cramér continues by recommending the use of confidence intervals as estimates of measurement uncertainty:

$$P(c_1 < \varepsilon < c_2; \varepsilon) = 1 - \alpha$$

> The probability that such and such limits (c_1, c_2 which may vary from sample to sample) include between them the parameter value ε (an unknown "measurement error") corresponding to the actual sample, is equal to $P_{success} = 1 - \alpha$ (the "risk of error").

Apart from limited-number statistics, uncertainties will also be determined by how well the ability of a particular cohort member matches the item task difficulty. This is because, as we described in relation to Fig. 11.4, counted-fraction data is only sensitive in mid-range, so that cohort members who are not challenged or who are too challenged by the task will have larger uncertainties compared cohort members with abilities matching the task difficulty.

To provide a representative sampling, adding analyses with additional items can be expected to improve reliability, not only by increasing the number of degrees of freedom, but also by varying the level of difficulty by choosing a number of different items which together span the range of interest. As stressed in the section above, 11.1.3 *A*: *Entity attribute description and specification*, items can be arranged in a hierarchy of item difficulties: choosing items of greater difficulty will challenge the healthier cohort members, while easier tasks enable a fair assessment of the less healthy.

Wright [97] described how the RMT makes separate estimates of attributes of each test person (TP) *i* with attribute (e.g., ability) θ_i and of each item *j* with attribute (e.g., difficulty) δ_j. These two parameters are adjusted in a logistic regression of Eq. (11.1): $\theta - \delta = \log\left(\frac{P_{success}}{1-P_{success}}\right)$, to the score response data $y_{i,\ j}$ on an ordered category scale by minimising the sum of the squared differences:

$$\sum_{i=1}^{N_{TP}} \sum_{j=1}^{L} \left(y_{i,j} - P_{success,i,j}\right)^2$$

The goodness of fit can be judged by examining how closely the overall fitted ogive item response curve of Eq. (11.1) matches individual average scores at different locations across the scale [6].

Continuing the discussion of the distribution variance in the Rasch [83] model, estimates of measurement uncertainty, u, for person (i, N_{TP}) and item (j, L) attributes and categories *k, k'*, derived from the Rasch expression Eq. (11.1) are made, respectively with the following expressions ([97] and Fig. 11.2):

$$\begin{cases} u(\theta) = \sum\limits_{j=1}^{L} \left(P_{success,i,j,k} \cdot P_{success,i,j,k\prime}\right)^{-\frac{1}{2}} \\ u(\delta) = \sum\limits_{i=1}^{N_{TP}} \left(P_{success,i,j,k} \cdot P_{success,i,j,k\prime}\right)^{-\frac{1}{2}} \end{cases} \tag{11.4}$$

The dichotomous relations of the basic Rasch model can be extended to the polytomous case, according to the expression:

$$q_{i,j,c} = P(y_{i,j} = c) = \frac{e^{\left[c\cdot(\theta_i - \delta_j) - \sum\limits_{k=1}^{c} \tau_{k,j}\right]}}{\sum_{c=0}^{K_j} e^{\left[c\cdot(\theta_i - \delta_j) - \sum\limits_{k=1}^{c} \tau_{k,j}\right]}} \tag{11.5}$$

The polytomous Rasch variant of GLM then models the response Y at any one point on the scale as a sum of dichotomous functions expressed as the log-odds ratio $z = \theta - \delta + \tau$ for each threshold τ, where the latter is where there is 50% probability that the response scores in either of the two adjacent categories. This polytomous Rasch variant is referred in the literature to the Andrich [5] "rating scale", Masters [56] "partial credit" approaches, amongst others. In R this is referred to as Extended Rasch Modeling: The R Package eRm. Programs such as WINSTEPS [96] and RUMM make a logistic regression of the polytomous Rasch formula to the response data $\mathrm{Y} = P_{success}$, using the "Joint Maximum Likelihood Estimation" method to estimate values of the 'latent' (explanatory or covariate) variables Z: θ, δ and the thresholds τ.

Naturally, the requirements of measurement – such as about invariance and dependency – need to be tested quantitatively in any specific application of a model (Eq. 11.1) to a set of data. Linacre & Fisher [49] write:

> An advantage of Rasch methodology is that detailed analysis of Rasch residuals provides a means whereby subtle inter-item dependencies can be investigated. If inter-item dependencies are so strong that they are noticeably biasing the measures, then Rasch methodology supports various remedies.

The statistical uncertainties (evaluated with Type A methods [42]) associated with the rate of success (Eq. 11.4), can be described with the Poisson distribution in Rasch's original [83] model (and associated with the response of the measurement system (Fig. 11.1b). Over and above type A evaluations, a complete expression of measurement uncertainties will also include accounts of a lack of knowledge about other elements of the measurement system, particularly the measured object (e.g., task) and the measurement instrument (person). In the same way as a Rasch approach insists on separate analysis of task difficulty and person ability, even measurement uncertainties are best evaluated in terms these elements separately. Examples of such uncertainties arise, as mentioned earlier, for instance when either the task differs – for one reason or another – from the "standard" recipe, perhaps arising from varying ways of administering the test (section above 11.3.1.1 *Reference measurement procedures: Construct specification equations as recipes for traceability*) or an individual rates his response on another scale than the cohort average, perhaps because of special sensitivities arising from emotion associated with stigma or illness (section 11.3.2.3 *Discrimination*).

Once again, when regarding measurement uncertainties not merely as standard deviations, but rather as estimates of "unknown measurement errors" [73], an MSA approach is appropriate when capturing the causes of uncertainty arising from less than perfect measurement systems. There are a number of similarities with the corresponding historic lack of attention to the "instrumental task of quantification" as mentioned in the section 11.2.3 *A way forward for measuring PCOs: Human as a B: Measurement Instrument*. Both measurement uncertainty analyses in general as well psychometric studies benefit from an MSA-based methodology.

Uncertainties in both the measured object (task difficulty) and measurement instrument (person ability) will propagate through the measurement system as corresponding uncertainties in the raw response data (and by extension to measurand estimation on restitution). An important part of a detailed analysis of Rasch residuals is thus a proper account of the peculiar sensitivity of the instrument (person) in a person-centered measurement system. Recall that "sensitivity" is the response of an instrument to a given input stimulus. Because of the strongly "resonant-like" variation of sensitivity as a function of attribute level, uncertainties in object or instrument at either extreme end of the scale will contribute essentially nothing to response uncertainties, while uncertainties at attribute levels closer to mid-scale will result in positive and negative swings in fit residuals on either side (as explained in relation to Fig. 11.4). A full account, including analytical expressions for construct alleys for various scale distortions, as well as the effects of discrimination can be found in section 5.7 of [75].

11.3.2.3 Discrimination

Referring again to Fig. 11.3, showing hierarchies of metrological standards, a requirement of the basic RMT is that person ability can be determined irrespective of task difficulty (and *vice versa*). That requirement is never fully satisfied, and we can envisage special cases, particularly in PCOs, where we might fail to meet it quite dramatically. When regarding a human being as an "instrument", again, with all the complexities that human behavior implies, one can consider cases where an 'irrational' response of a test person – caused by emotions or illness, for instance – could occur for one or more tasks in which the scaling of person ability for a given task difficulty deviates from the responses overall for the cohort. It is, so to say, a differential item and person response which leads to scale distortions which are not automatically compensated for with the basic RMT, but require an additional parameter, such as person discrimination (i.e., instrument sensitivity). Examples include [63]:

- Acquiescence: "Agreement regardless of item content"
- Disacquiesence: "Disagreement regardless of item content"
- Extreme response bias: "Use scale endpoints regardless of item content"
- Middle response bias: "Use scale midpoint regardless of item content"
- Social desirability bias: "Present oneself in a positive way, regardless of item content"

Scale distortions associated with any of these effects would be additional to the counted fraction non-linearity modelled with RMT which has to be compensated for when making a proper analysis of raw score ordinal data Evidence for such effects can be sought in the residuals of fit and various plots, such as the so-called construct alleys ([54], section 5.7 of [75]). A construct alley, i.e., a plot of values of task difficulty values, δ (or person ability values, θ) against the residuals, such as *INFIT* – *ZSTD*, of the logistic regression, is a sparsely used but potentially powerful tool to diagnose response patterns and further enhance the understanding of fit statistics. Apart from random noise, Massof et al. [54] have reported systematic distortions in construct alleys for different vision traits (for tasks of mobility, reading, visual information and visual motor). Another tool for analyzing goodness of fit of the logistic Eq. 11.1 found to be useful in detecting scale distortions is the principal component loading plot [78].

The form of the ogive response curve exemplified in Fig. 11.4, particularly how fast the ogive curve rises through the mid-point, is in general determined by measurement uncertainty associated with how sensitive each instrument (test person) is to the task at hand [60].

A basic requirement of Rasch's models for measurement is that every test person lies on the same ogive curve, irrespective of who they are. Apart from task difficulty and instrument ability, in some cases an additional factor – the discrimination of the person responding – may vary (for instance, because of illness or emotion [86]). This can be captured in related Item Response Theory (IRT) with a discrimination parameter: the finite resolution, ρ, of the instrument (a patient in the present case) modelled as the entropy $H(Z, Y) \sim \ln(\rho) = \ln(\sqrt{3} \cdot 2 \cdot u)$ of a uniform distribution associated with limitations in measurement quality as measurement information is transmitted from the object attribute, Z, to the response, Y, of the instrument, where u is the standard measurement uncertainty [77].

There are, of course, different conceivable models of a modified response associated with effects such as acquiescence and response bias at various locations of the scale (section 5.7 of [75]). A simple rescaling, centered on the logistic scale and varying linearly would be $\partial\delta_j = s \cdot \delta_j$, where s is a re-scaling factor. For some reason or other, rating of item *j* is made so that the item attribute (such as task difficulty) lies on a different scale: a positive value τ of rescaling indicates an extended scale (test persons (instruments) rate this item more strongly than others, perhaps to indicate an increased importance or weight), while a negative value τ of rescaling corresponds to the case where rating does not recognize a reversed scale. An example of the latter is where a survey designer has deliberately included alternately positive (true key) and negative (false key) items, perhaps to reveal evidence of acquiescence in raters, that is, responses which tend to agree with questions (e.g., personality scales [87]) without due regard for the content of the item.

11.3.2.4 Measurement Uncertainty and Measurement System Analysis

Summarizing the steps to be taken overall when expressing measurement uncertainty [42]:

1. Analyze the measurement system. Set up an error budget
2. Correct for known measurement errors, such as subtracting for known bias, *b*, and sensitivity *K* differing from 1 (unity) (Fig. 11.3)
3. Evaluate (standard) measurement uncertainties with methods of type A (i.e., statistically) alternatively type B
4. Combine standard measurement uncertainties by quadratic addition $= > u_c$
5. Expand measurement uncertainty $= > U = k \cdot u_c$

The first (and perhaps most essential) step in any evaluation of measurement uncertainty is to make so complete, valid and reliable description of the measurement or classification system at hand. Analogously to the corresponding first step on construct description, if one misses a key contribution when formulating one's measurement model, then no amount of statistics or other actions in the later steps of the list above will compensate for that omission.

The techniques of MSA are recommended, not only when ensuring metrological traceability, but also here where there are preferably clear and separate estimates of the contributions of each element of the measurement system – instrument, operator, environment and measurement method (Fig. 11.1), using methods such as ANOVA – to the overall uncertainties in the response when measuring a certain object.

Further discussion can be found in Chapter 5 of [75].

11.4 Decision Risks and Uncertainty

In this final section, we complete the measurement task by closing the quality loop, by returning to where we started and complete quality-assurance of a measurement task by considering how best to make a final assessment of conformity of the entity to specified requirements in terms of impact. As said at the outset, most measurements are not ends in themselves, but are usually made with the aim of assessing product – in the present case the health care service provided in terms of quality of PCC. When closing the loop and providing the final "bookend", several of the quality characteristics of health care services identified at the start of this chapter will be key to assuring quality in PCC.

At this final stage, use can be made of all the previous material presented in the preceding sections of this chapter.

11.4.1 Comparing Test Result with Product Requirement

A test result for the cognitive ability, θ, of a test person, with its measurement uncertainty interval, compared with examples of the lower specification limits $L_{SL,\ AD}$ and $L_{SL,\ MCI}$ for diagnosis by a clinician of Alzheimer's disease (AD) and mild cognitive impairment (MCI), respectively is shown in Fig. 11.5 as a typical case of PCC (data taken from Melin & Pendrill [61]). A lower specification limit means that measurement values less than the limit are judged to be a positive indication of disease, so that for instance test results of ability, θ, less than $L_{SL,\ AD}$are considered to lie probably in the Alzheimer's disease region of permissible values R_{AD}.

The specification limits and corresponding regions of permissible values in the case shown in Fig. 11.5 have been set according to the 50% and 35% limits for 'probable' and 'possible' AD set by Hughes et al. [32]. It should however be noticed that such specification limits vary in clinic and assessments of memory ability is not the only input to the clinical examination to set diagnosis.

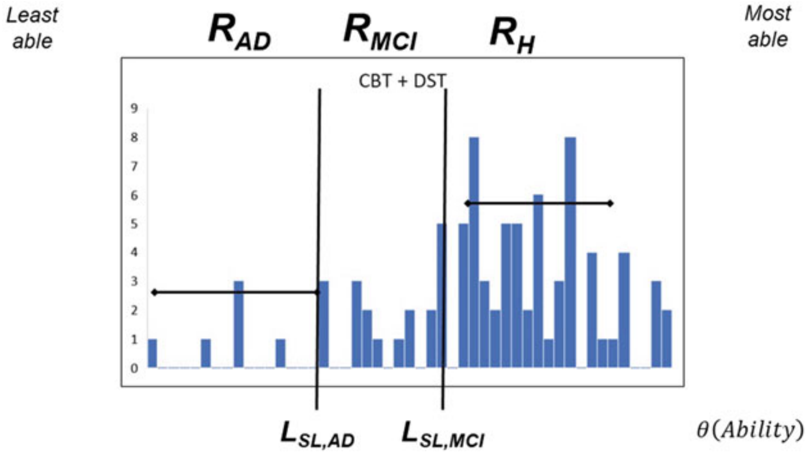


Fig. 11.5 CBT and DST memory test results [62], with measurement uncertainty intervals (double-ended arrow, $k = 2$), together with specification limits for diagnosis of Alzheimer's disease (AD) and mild cognitive impairment (MCI) Hughes et al. [32]. The histogram columns indicate the distribution of person ability across the cohort

11.4.1.1 Requirements for Decision Risk Management

Measurement uncertainty in a test result – an apparent product dispersion arising from limited measurement quality (section 11.3.2.2 *Errors and uncertainty. Reliability*) – can be a concern in conformity assessment by inspection since, if not accounted for, uncertainty can both lead to incorrect estimates of the consequences of entity error and to an increase in the risk of making incorrect decisions, such as failing a conforming entity or passing a non-conforming entity when the test result is close to a tolerance limit (section 11.4.2.2 *Consumer and Provider Risks*).

Requirements for appropriately accounting for the consequences of measurement uncertainty when making decisions of conformity have recently entered more prominently in the main written standards for conformity assessment, such as the latest version of ISO 17025:2017 [39], which states:

> 7.7.1 Evaluation of conformance
>
> When statement of conformity to a specification or standard for test or calibration is requested, the laboratory shall:
>
> document the decision rules employed taking into account the level of risk associated with the decision rule employed (false accept and false reject and statistical assumptions associated with the decision rule employed);
>
> apply the decision rule.
>
> NOTE For further information see ISO/IEC Guide 98-4. [43]

For PCOs addressing cognition, for example, (section 11.1.5.2 *PCOs: Example neuropsychological cases*) the “laboratory” referred to be above could be a memory clinic. For patient participation (section 11.1.5.3 *PCOs: Example patient participation*), the “laboratory” could correspond to the survey administrator such as authorities, clinic lead or research group.

Because of measurement uncertainty (as evaluated in the section 11.3.2 *Errors and uncertainties in PCOs*) or as otherwise arising when making a clinical decision), there are risks of making incorrect decisions of conformity where a test result can appear to lie in one region, but there is a finite probability of lying in another region, particularly where the test result is close (within uncertainty) of the decision limit (section 11.4.2.2 *Consumer and Provider Risks*). A decision of conformity of a quality characteristic with respect to a specification limit, normally made in the present case by a clinician, but in some cases as a self-assessment by the patient herself, depends on two factors: the measurement uncertainty and the distance between a test result and the specification limit. These two factors together determine if an entity is approved or not and what the risks of incorrect decision are (section 11.4.2.2 *Consumer and Provider Risks*).

A definition of decision-making accuracy – in terms of $P_{success}$, that is the probability of making a correct classification – comes from the expression (Chapter 2 in [75], Fig. 11.2):

$$\begin{aligned} Accuracy\,(decisionmaking) = {} & response\ categorization \\ & - input\,(true)\ categorization \end{aligned} \tag{11.7}$$

RMT is very broad in its application and is not restricted to measurement systems with human intervention but should also apply to other 'probe: task' systems. Examples include [71–73, 92] the performance of a system (characterized by the ability of providing healthcare, such as waiting times for surgery, separately from the levels of challenge associated with each task), or the determination of material testing (e.g. the ability of an indenter, separately from the hardness of each material test block).

Apart from illustrating product conformity, plots such as Fig. 11.5 can also be used when considering *measurement* conformity. To make decisions about conformity to "product" specifications in a reliable manner will require measurements which themselves have been shown to satisfy conformity to corresponding measurement specifications. A measurement conformity assessment version of Fig. 11.5 could show for example how the actual instrument error (with its uncertainty interval) lies with respect to the maximum permissible (measurement) error (MPE) and maximum permissible (measurement) uncertainty (MPU) specifications.

11.4.2 C: Man as an Operator: Rating the Rater

In terms of measurement system analysis where a human enters into different parts of a measurement system (Fig. 11.1), the case of a clinician making a diagnosis on the basis of the data exemplified in Fig. 11.5, from the chapter by Melin and Pendrill [61], can be interpreted as a human acting as the Operator of the measurement system, as shown in Fig. 11.6 (*C* at the third main stage in the measurement process (Figure 1 of [77])). That is, the operator (clinician) makes a diagnosis, based on the response of the instrument (each test person), about whether that person's ability lies within or outside specification limits in Fig. 11.5 (corresponding to whether a product is "approved" or not in regular conformity assessments is "approved" or not).

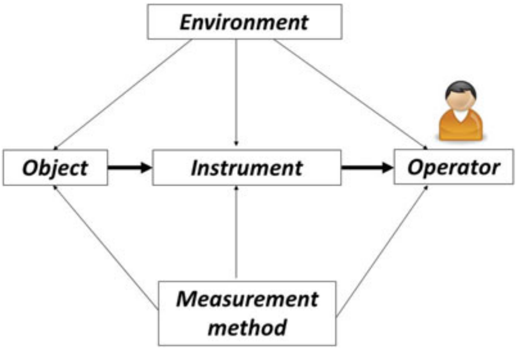


Fig. 11.6 A human as an operator in a measurement system

In recent work, Andrich [6] has studied Gaussian and Rasch distributions in instrument response, while van der Bles, et al. [13] considered how epistemic uncertainty is interpreted by the rater, as well as communication about uncertainty to a third-party audience. Our approach is instead to propose that RMT can usefully be applied to extend the 'rating the raters' approach of Akkerhuis et al. [3, 4] who have modelled decisions about manufactured product made in an industrial context. With our extension (chapter 6 of [75]), separate estimates of the ability, θ_i, of each rater, *i*, to classify product can be made, alongside estimates of the level of difficulty, δ_j, of each product classification task.

But first, in the next section (11.4.2.1), we need to identify the main constructs with which a rater – such as a health care professional – can be characterized for conformity assessment in quality assurance, for instance in PCC.

11.4.2.1 Rater Constructs

A comprehensive description of the constructs of interest when making clinical diagnoses can be based on the preceding techniques and construct specification equations (CSE) can be formed, in analogous fashion (section 11.3.1.1 *Reference measurement procedures: Construct specification equations as recipes for traceability*). Apart from how well trained, how experienced, and how alert a health care professional is, their ability θ to diagnose may also reflect their attitudes to desiring to give high quality care and their ability to "leave prejudices at the door" when meeting a new cohort individual [66]. Similarly, explaining what makes the clinical classification of different cohort individuals more or less difficult, level δ, probably depends in a complex way on how ill each assessed person is. It is unclear whether it is easier or more difficult to diagnose a healthy or sick person.

With the overall aim of supporting interoperability in the exchange of meaningful information between information systems in respect of nursing diagnoses and nursing actions, the standard ISO 18104 specifies **healthcare entities for nursing diagnoses**. Motivation for the development of terminological systems to support nursing in ISO 18104 [40] lies in multiple factors including the need to describe nursing in order to educate and inform students and others, represent nursing concepts in electronic systems and communications, including systems that support multiprofessional team communications and personal health records, and analyses data about the nursing contribution to patient care and outcomes – for quality improvement, research, management, reimbursement, policy and other purposes.

Connecting the terminology of ISO 18104 [40] in PCC (nurses below can be any clinical profession):

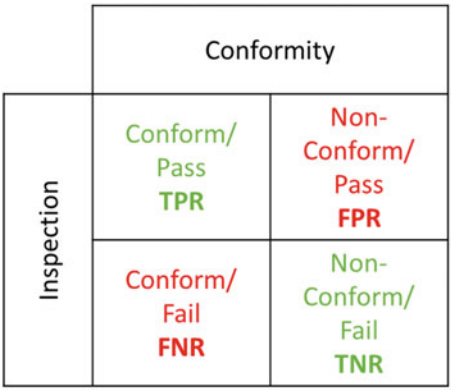
- The probability $P_{success,\ i}$ of a successful clinical diagnosis by an <individual> nurse on a <<recipient of care>> would correspond to a measure of how well a <judgement> is performed (section 11.4.2.2 *Consumer and Provider Risks*).
- The <<recipient of care>> is the person, family, group, or other aggregate to whom the action is delivered.

- The entity (object) is the <<target>> of the diagnosis which is the entity that is affected by the nursing action or that provides the content of the nursing action. Semantic categories in the <<target>> domain include but are not limited to: <body component>, <sign>, <device>, <substance>, <physical environment>, <resource>, <process>, <dimension>, <individual>, <group>, and the categories that have the role of <<focus>> in nursing diagnoses. Nursing diagnosis can also be a <target>.
- A construct is a <focus> which is an "area of attention", such as "tissue integrity, body temperature, activity of daily living". <<Focus>> may be qualified by <timing>.
- A <judgement> (opinion or discernment related to a <focus>) may be characterized as being "Impaired, reduced, ineffective".
- Judgement categories that are valid for representation of a nursing diagnosis include, but are not limited to, alteration, adequacy, and effectiveness.
- Attributes associated with nursing performance (such as knowledge, motivation, and ability) belong to <dimension>.

11.4.2.2 Consumer and Provider Risks

Four alternative outcomes – illustrated in Fig. 11.7 – can be encountered when making decisions of conformity in the presence of measurement uncertainty. In addition to a pair of correct decisions of conformity, measurement uncertainty can lead to a second pair:

- non-conforming entities being incorrectly passed on inspection – consumer risk, α, that is, in PCC, the person receiving care
- correctly conforming entities being incorrectly failed on inspection – provider risk, β, that is, in PCC, the care provider



	Conformity	
Inspection	Conform/ Pass **TPR**	Non- Conform/ Pass **FPR**
	Conform/ Fail **FNR**	Non- Conform/ Fail **TNR**

Fig. 11.7 Two correct & two incorrect decisions of compliance. *TPR* = True Positive Rate, *FPR* = False Positive Rate, *FNR* = False Negative Rate and *TNR* = True Negative Rate

Particularly when a test result is close (within the uncertainty interval) to a specification limit [73, 75]. Making decisions in the presence of measurement uncertainty are an example of the wider concept of classification.

The risks of incorrect decisions on identification [41] – both by variable and by attribute – are simply connected to the logistic regression formula (Eq. (11.1)) by the relation:

$$1 - \alpha = P_{success}$$

Making decisions for a single item – in the present case, an individual patient ("consumer") receiving PCC from a care provider – the specific risks can be readily calculated in terms of the area under that portion of the measurement PDF g_{test} which lies in each case, so to say, "beyond – on the other side of" the relevant specification limit, L_{SL}, to the mean $\bar{z}$, and are expressed mathematically as:

Consumer specific risk (by variable)

$$\alpha_{specific}(\bar{z}) = \int_{\eta < L_{SL}} g_{test}(\eta|\bar{z}) \cdot d\eta(\bar{z} \geq L_{SL}) \tag{11.8}$$

Provider specific risk (by variable)

$$\beta_{specific}(\bar{z}) = \int_{\eta \geq L_{SL}} g_{test}(\eta|\bar{z}) \cdot d\eta(\bar{z} < L_{SL})$$

– for a test result mean value, $\bar{z}$, (distribution g_{test}) and a lower specification limit, L_{SL}.

These decision risks can be summarized in a so-called confusion matrix:

$$\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

which corresponds to the four quadrants shown in Fig. 11.7.Apart from specific risks when assessing conformity to specification of a single item (measurement distribution g_{test}), in general different items (occasions of PCC) will have different quantity values – for instance, due to actual variations (distribution g_{entity}) in health of different patients (see section above 11.2 *Benefits of combining Rasch Measurement Theory (RMT) and quality assurance*) – and the commensurate global risks of incorrect decisions also need to be estimated. Extensions of the formulae above by convoluting the variables Z_{global} and Y_{global} leads to expressions such as:

Consumer global risk (by variable)

$$\alpha_{global}(\bar{z}) = \int\int_{\eta < L_{SL}} g_{entity}(\xi|\bar{z}) \cdot g_{test}(\eta|\bar{z}) \cdot d\eta \cdot d\xi(\bar{z} \geq L_{SL})$$

Risks and the consequences of incorrect decision-making in conformity assessment should always be evaluated. Beyond the percentage probabilities discussed in this section, ultimately risks can be minimized by proactively setting limits on maximum permissible measurement uncertainties and on maximum permissible consequence costs [73].

11.4.3 Receiver Operating Characteristics: A Human as an Operator & Rating the Rater

For historical reasons, there are a number of related plots under the name "operating characteristic" which intend to convey an indication of the "power" of any measurements:

In statistical acceptance sampling, the "operating characteristic" (or "discriminatory power") curve is a plot of the probability of accepting a lot as a function of an explanatory variable [68]

A "receiver (or "relative") operating characteristic" (ROC) can be a plot of the true positive rate (*TPR*, or "sensitivity") [81, 98]:

$$TPR = \frac{TP}{TP + FN} = \frac{1 - \alpha}{1 - \alpha + \beta} \tag{11.9}$$

against the false positive rate (*FPR*), where β is the "supplier" risk in a dichotomous case, i.e., the probability that product is incorrectly rejected ('false negative', *FN*).

Two possible methods to plotting ROC curves for calculating the probability of incorrect decisions α and β:

using Eq. (11.8) based on uncertainties, e.g., from the Rasch analyses; or
simply counting the rate of correct clinical diagnosis at each signal level.

Potential differences between these approaches might arise if the clinician uses other techniques, in addition to inspection of plots such as Fig. 11.5. That is the most likely case since the clinician will typically weigh together several factors (including his or her professional experience) when reaching a final decision. The clinical diagnosis uncertainties in method (ii) are probably greater than the signal level uncertainties of method (i) because the clinical judgment includes a number of criteria, in addition to signal levels. That this is so in the present case is evident for the data shown in Fig. 11.8, since cohort individuals are diagnosed as sick across practically the whole range of signal levels. Further consideration of possible explanations of clinician ability are given in the section 11.4.3.1 *Explaining clinical diagnoses*.

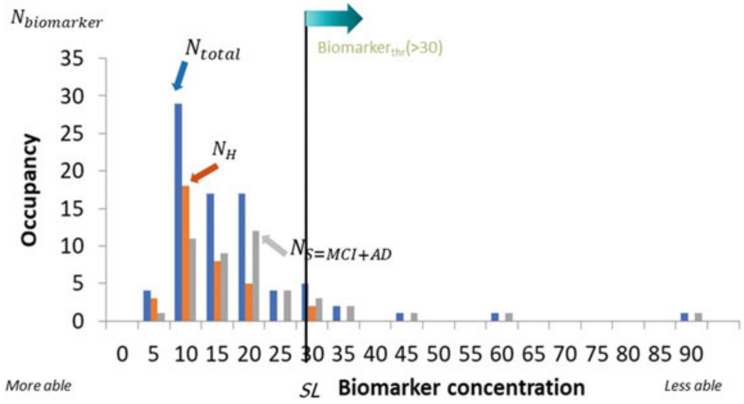


Fig. 11.8 Example of basis for calculating ROC curves in a biomarker case. Histograms of the occupancy (number of cohort members, $N_{biomarker} = 80$) as a function of measured biomarker concentration (arbitrary units). Clinical (binary) classification $H = healthy$ and $S = sick = MCI + AD$ indicated by different colored columns: Blue $\boldsymbol{N_{total} = N_H + N_S}$; Orange $\boldsymbol{N_H}$, number of healthy cohort individuals, and Grey $\boldsymbol{N_S}$, number of unwell cohort individuals

Calculating ROC curves with method (ii) on the basis indicated in Fig. 11.8 (using data from the example provided by Melin & Pendrill [61]) involves evaluation of expressions such as [81, 98]:

$$Sensitivity = TPR = \frac{\sum_{C>SL} N_S}{\sum_C N_S} \tag{11.10}$$

$$False\ positive\ rate = FPR = \frac{\sum_{C>SL} N_H}{\sum_C N_H}$$

where clinical classification is made of the occupancy, the number N of cohort individuals, in one of two categories $H = healthy$ and $S = sick = MCI + AD$, at each signal level, C, (in the present case, the concentration of a selected biomarker). Both the true positive rates TPR and the rate of false positive rates FPR (as well as the respective expressions for negatives) are calculated in terms of the number of cohort individuals having signals, C, exceeding a signal specification limit, SL, which is varied successively over the signal range of interest, assuming that higher signal levels should indicate a greater probability of sickness.The relative performance ("discriminatory power") of different signals in providing a base for clinical diagnosis can be gauged by comparing ROC curves for the different signals. In addition to biomarker concentration, one can also evaluate Eq. 11.10 using signals from Rasch cognitive ability θ, for instance.

The resulting ROC curves – illustrated for two different signals in Fig. 11.8 (example provided by Melin & Pendrill [61]) – are obtained by plotting Sensitivity TPR versus False positive rate FPR over the range of each signal level (C) of interest.

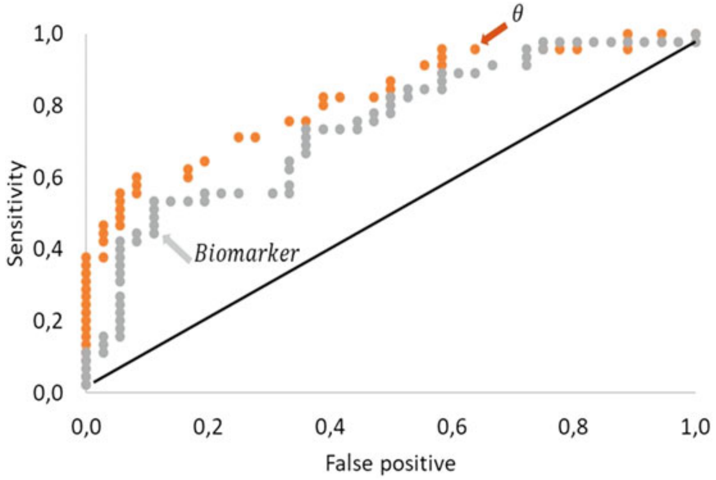


Fig. 11.9 Two ROC curves of Sensitivity ***TPR*** versus False positive rate ***FPR*** based on clinical classification over the range of each signal level (C) of interest: Orange dots, Rasch cognitive ability ***θ***; Grey dots, biomarker concentration

The ideal ROC curve in terms of greatest discriminatory power is the one furthest to the left-hand upper corner [81].

Each ROC curve is obtained by calculation of Eq. 11.10 by varying the signal specification limit, *SL*, from low signal levels (*C*) – top-right of each ROC curve – where all decisions are positive – to increasingly higher signal values, causing both decreased sensitivity and the rate of false positives to reduce, reaching minimum values at bottom-left of each ROC curve. The discriminatory power of each signal is revealed by how well each kind of signal can maintain the highest sensitivity as signal levels increase. A traditional figure of merit is the Area Under the Curve (AUC), found by integrating the ROC curve across the range of FPR [81]. (Such a figure of merit may however suffer from the uncorrected effects of counted-fraction ordinality – see the section 11.1.5.1 *PCC including physical, psychological and social integrity*.)

11.4.3.1 Explaining Clinical Diagnoses

To approach as closely as possible the same diagnosis criteria used by clinicians, the signal chosen for ROC curve analysis should be the most representative, comprehensive and valid.

In reflecting over what lies behind making certain signals better than others to diagnose correctly, as revealed in Fig. 11.8, one can conjecture that a composite memory measure is arguably a more comprehensive measure, and perhaps "closer" to the clinician's assessment method, than say biomarker concentration when making a clinical classification between healthy and unwell cohort individuals. As in this case, a diagnosis of Dementia due to AD is not based only on either memory ability of biomarker concentrations. Rather it comprises criteria such as the ability to function at work or at usual activities and mental status examination or neuropsychological testing together with biomarkers.

Arguably a composite memory measure (naturally Rasch-transformed, section 11.1.5.1 *PCC including physical, psychological and social integrity*) – preferably with a fully developed construct specification including all significant explanatory variables [61] – is likely to be a better choice than, say an individual biomarker, particularly in the context of PCC.

11.4.4 Many-Body Modelling and Conclusions

In this chapter we have considered measurement models where either the patient (Fig. 11.1b) or the clinician (Fig. 11.6) act solely when assuring quality of PCOs, from start to finish of the quality loop.

Bringing the patient "into a partnership with health care professionals" as a key aspect of PCC (section 11.1.2 *Quality assurance in person-centered care. Design of experiments*) opens up a new field of quality-assured measurement suitable for future studies. To model the patient: clinician partnership one could envisage extending the "single-body" basic Rasch model to a "many-body" version. Such a modelling has indeed already been done, where two people of different abilities jointly tackle a task of a given difficulty, such as in the game of chess. Brinkhuis and Maris [15] consider how the famous chess rating system Elo can be applied in education to student monitoring. Such models can be extended to any number of mutually interacting people, for instance if one wishes to study social interactions in an area called "social physics" [79]. Such work would extend our study of how informational entropy can be used to model the ability (or more generally attitude) of an individual [61] or the ability of an organization [99] to many-body situations, such as the patient: clinician partnership in PCC.

Acknowledgments This chapter is based on learnings and outcomes in several projects during the past years, including the VINNOVA funded project Quality assured metrology in health care (Dnr, 2017-05011), partnership with Region Skåne and Innovation Skåne, as well as the RISE internal platform for Categorical Based Measures. Hence, we would like to express our great appreciation to collaborators and partners for valuable and constructive work together. Part of the work reported in this chapter has also been part of the 15HLT04 NeuroMET project which received funding from the EMPIR programme co-financed by the Participating States and from the European Union's Horizon 2020 research and innovation programme. We would therefore offer our special thanks to partners at Charité University Hospital for patent recruitment and memory test data as well as Physikalisk Technische Bundesanstalt (PTB), CHU Montpellier and LGC UK National Measurement Laboratory for analyses of brain volumes and biomarkers.

References

1. AIAG, Measurement systems analysis reference manual, in *Chrysler, Ford, General Motors Supplier Quality Requirements Task Force*, (Automotive Industry Action Group, 2002)
2. J. Aitchison, The statistical analysis of compositional data. J. R. I. State Dent. Soc. **44**, 139–177 (1982)
3. T. Akkerhuis, *Measurement System Analysis for Binary Tests* (PhD Thesis, FEB: Amsterdam Business School Research Institute (ABS-RI), 2016) ISBN 9789462333673
4. T. Akkerhuis, J. de Mast, T. Erdmann, The statistical evaluation of binary test without gold standard: Robustness of latent variable approaches. Measurement **95**, 473–479 (2017)
5. D. Andrich, A rating formulation for ordered response categories. Psychometrika **43**, 561–573 (1978)
6. D. Andrich, *On an Identity Between the Gaussian and Rasch Measurement Error Distributions: Making the Role of the Instrument Explicit* (IOP Conf. Series: J. Phys.: Conf. Series 1065 072001, 2018)
7. ASTM E2782 - 11:2011, *Standard Guide for Measurement Systems Analysis (MSA)* (ASTM, 2011)
8. S.-V. Bala, K. Forslind, B. Fridlund, P. Hagell, Measuring person-centered care in nurse-led outpatient rheumatology clinics. Musculoskeletal Care **16**, 296–304 (2018). https://doi.org/10.1002/msc.1234
9. E. Bashkansky, T. Gadrich, I. Kuselman, Interlaboratory comparison of test results of an ordinal or nominal binary property: analysis of variation, Accred Qual Assur 17:239–243, (2012). https://doi.org/10.1007/s00769-011-0856-0
10. J.P. Bentley, *Principles of Measurement Systems*, 4th edn. (Pearson Education Limited, London, 2005)
11. D.M. Berwick, What 'patient-centered' should mean: Confessions of an extremist. Health Aff. **28**, w555–w556 (2009)
12. M.J. Bitner, B.H. Booms, M. Stanfield Trereault, The service encounter: Diagnosing favorable and unfavorable incidents. J. Mark. **54**, 71–84 (1990)
13. A.M. van der Bles, S. van der Linden, A.L.J. Freeman, J. Mitchell, A.B. Galvao, L. Zaval, D.J. Speigelhalter, Communicating uncertainty about facts, numbers and science. R. Soc. Open Sci. **6**, 181870 (2019)
14. M. Boers, J.R Kirwan, P Tugwell, et al., The OMERACT Handbook, Published by OMERACT (2018)
15. M.J.S. Brinkhuis, G. Maris, *Dynamic Parameter Estimation in Student Monitoring Systems* (Measurement and Research Department Reports 2009–1, Cito, Dutch National Institute of Educational Measurement, 2009)
16. S. Cano, T. Vosk, L.R. Pendrill, A. Stenner, On trial: The compatibility of measurement in the physical and social sciences. J. Phys. Conf. Ser. **772**, 012025 (2016)
17. S.J. Cano, L.R. Pendrill, S.P. Barbic, W.P. Fisher, Patient-centered outcome metrology for healthcare decision-making. J. Phys. Conf. Ser (2017)
18. H. Cramér, in *Mathematical Methods of Statistics, Princeton Mathematical Series*, ed. by M. Morse, H. Robertson, A. Tucker, (Hugo Gebers Förlag, Almqvist & Wiksell, 1945) ISBN 9780691005478
19. J. Dagman, R. Emardson, S. Kanerva, L.R. Pendrill, A. Farbrot, S. Abbas, A. Nihlstrand, *Measuring comfort for heavily-incontinent patients assisted by Absorbent products in several contexts* (*IMECHE*, London, 2013) 5–6 November 2013
20. D de Silva, Helping measure person-centred care, Evidence review, The Evidence Centre (UK), (2014). https://www.health.org.uk/publications/helping-measure-person-centred-care
21. W.E. Deming, *Out of the Crisis* (Massachusetts Institute of Technology, Center for Advanced Engineering Study, Cambridge, MA, 1986) ISBN 0911379010. OCLC 13126265
22. R. Dybkaer, *An ontology on property for physical, chemical, and biological systems*, ISBN 978-87-990010-1-9, (2009). http://ontology.iupac.org

23. D. Edvardsson, A. Innes. Measuring Person-centered Care: A Critical Comparative Review of Published Tools. The Gerontologist, **50**(6), 834–846, (2010). https://doi.org/10.1093/geront/gnq047
24. Ekman et al., Person-centered care – Ready for prime time. Eur. J. Cardiovasc. Nurs. **10**, 248–251 (2011)
25. EN 15224:2012 and 2017, *Quality management systems – EN 9001:2015 for healthcare*
26. W.P. Fisher Jr., Physical disability construct convergence across instruments: Towards a universal metric. J. Outcome Meas. **1**(2), 87–113 (1997)
27. R. Fleischmann, Einheiteninvariante Gröβengleichungen, Dimension. Der Mathematische und Naturwissenschaftliche Unterricht **12**, 386–399 (1960)
28. D.D. Gremler, The critical incident technique in service research. J. Serv. Res. **7**, 65–89 (2004)
29. M. Hallworth et al., Med Clin Chem **61**, 589–599 (2015)
30. E. Harding, S. Wait, J. Scrutton, The state of play in person-centered care: A pragmatic review of how person-centered care is defined, applied and measured, featuring selected key contributors and case studies across the field. The Health Policy Partnership (2015) Retrieved from: http://www.healthpolicypartnership.com/wp-content/uploads/State-of-play-in-person-centered-care-full-report-Dec-11-2015.pdf
31. HSL 2017: 30, TITLE II. Provisions for All Health and Health Care Health and Medical Services Act (SFS 2017:30) [Swedish: AVDELNING II. BESTÄMMELSER FÖR ALL HÄLSO- OCH SJUKVÅRD Hälso- och sjukvårdslag (2017:30)]
32. L.F. Hughes, K. Perkins, B.D. Wright, H. Westrick, Using a Rasch scale to characterize the clinical features of patients with a clinical diagnosis of uncertain, probable, or possible Alzheimer disease at intake. J. Alzheimers Dis. **5**, 367–373 (2003)
33. S.M. Humphry, The role of the unit in physics and psychometrics. Meas Interdiscip Res Perspect **9**(1), 1–24 (2011)
34. R.J. Irwin, A psychophysical interpretation of Rasch's psychometric principle of specific objectivity. *Proceedings of Fechner Day*, 23 (2007)
35. ISO 2020, *What is conformity assessment? Health informatics – Categorial structure for terminological systems of surgical procedures*
36. ISO 10012:2003 *Measurement management systems -- Requirements for measurement processes and measuring equipment*, ISO
37. ISO 10576-1 *Statistical methods – Guidelines for the evaluation of conformity with specified requirements*
38. ISO/IEC 17000:2004, *Conformity Assessment – General Vocabulary*, International Organization for Standardization, Geneva
39. ISO/IEC 17025:2017 *General requirements for the competence of testing and calibration laboratories*,
40. ISO 18104:2014 *Health informatics – Categorial structures for representation of nursing diagnoses and nursing actions in terminological systems*,
41. G. Iverson, R. Luce, The representational measurement approach to psychophysical and judgmental problems, in *Measurement, Judgment, and Decision Making*, (Academic Press, Cambridge, 1998)
42. JCGM 100:2008, *Evaluation of measurement data – Guide to the expression of uncertainty in measurement (GUM 1995 with minor corrections)*, in Joint Committee on Guides in Metrology (JCGM)
43. JCGM 106:2012, *Evaluation of measurement data – The role of measurement uncertainty in Conformity Assessment*, in Joint Committee on Guides in Metrology (JCGM)
44. T. Jesus, F.A. Bright, C.S. Pinho, C. Papadimitriou, N.M. Kayes, C.A. Cott, *Scoping Review of the Person-Centered Literature in Adult Physical Rehabilitation* (Preprints 2019, 2016), p. 2019020015. https://doi.org/10.20944/preprints201902.0015.v1
45. M. Kaltoft, M. Cunich, G. Salkeld, J. Dowie, Assessing decision quality in patient-centered care requires a preference-sensitive measure. J. Health Serv. Res. Policy **19**, 110–117 (2014)

46. P. Kersten, P.J. White, A. Tennant, Is the pain visual analogue scale linear and responsive to change? An exploration using Rasch analysis. PLoS One **9**(6), e99485 (2014)
47. A. Leplege, F. Gzil, M. Cammellin, C. Lefeve, B. Pachoud, I. Ville, Person-centeredness: Conceptual and historical perspectives. Disabil. Rehabil. **29**, 1555–1565 (2007)
48. J.M. Linacre, Optimizing rating scale category effectiveness. J. Appl. Meas. **3**(1), 85–106 (2002)
49. J.M. Linacre, W.P. Fisher Jr., Harvey Goldstein's objections to Rasch measurement: A response from Linacre and Fisher. Rasch Meas Trans **26**(3), 1383–1389 (2012)
50. J.M. Linacre, B. Wright, The 'length' of a logit. Rasch Meas. Trans. **3**, 54–55 (1989)
51. J. Lindberg, M. Kreuter, L.O. Person, C. Taft, Patient participation in rehabilitation questionnaire (PPRQ) – development and psychometric evaluation. Spinal Cord **51**(11), 838–842 (2013)
52. K. Luhr, A.C. Eldh, U. Nilsson, M. Holmefur, Patient preferences for patient participation: Psychometric evaluation of the 4Ps tool in patients with chronic heart or lung disorders. Nord J Nurs Res. **38**, 68–76 (2017)
53. L. Mari, C. Narduzzi, G. Nordin, S. Trapmann, Foundations of uncertainty in evaluation of nominal properties. Measurement **152**, 107397 (2020)
54. R.W. Massof, L. Ahmadian, L.L. Grover, J.T. Deremeik, J.E. Goldstein, C. Rainer, C. Epstein, G.D. Barnett, The activity inventory: An adaptive visual function questionnaire. *Optom. Vis. Sci.* **84**, 763–774 (2007)
55. R.W. Massof, C. Bradley, *A strategy for measuring patient preferences to incorporate in benefit-risk assessment if new ophthalmic devices and procedures*, J Phys : Conf Ser. **772**, 012047 (2016) https://doi.org/10.1088/1742-6596/772/1/012047
56. G.N. Masters, A Rasch model for partial credit scoring. Psychometrika **47**, 149–174 (1982)
57. J.C. Maxwell, Molecules. Nature, 437–441 (1873a, 1873)
58. J.C. Maxwell, (1873b) Maxwell, James Clerk, 1873b, A treatise on electricity and magnetism, Clarendon Press, Oxford, [https://archive.org/details/atreatiseonelec03maxwgoog/page/n8]
59. B. McCormack et al., Person-Centeredness – The 'State' of the Art. *Int Pract Develop J* **5** (2015) Special Issue on Person-centeredness, Article 1
60. H. Melgaard, P. Thyregod, Acceptance Sampling by Variables under Measurement Uncertainty, in *Frontiers in Statistical Quality Control*, vol. **6**, (2001), pp. 47–57. ed. by H-L. Lenz, P-T. Wilrich, (Heidelberg; New York, Physica-Verl., ISBN 3-7908-1374-5)
61. J. Melin, L.R. Pendrill, The Role of Construct Specification Equations (CSE) and Entropy in the Measurement of Memory in Person-Centered Care, in *Person Centered Outcome Metrology*, ed. by S. Cano, P. Marquis, A. Regnault, W. Fisher, (Springer, 2022)
62. J. Melin, R. Fornazar, M. Spångfors, L.R. Pendrill, Rasch analysis of the patient participation in rehabilitation questionnaire (PPRQ). J. Eval. Clin. Pract. **26**, 248–255 (2020) https://doi-org.ezproxy.ub.gu.se/10.1111/jep.13134
63. G.J. Mellenbergh, *Counteracting Methodological Errors in Behavioral Research* (Springer Nature Switzerland AG, 2019) ISBN 978-3-319-74352-3
64. J. Michell, Measurement in psychology: A critical history of a methodological concept. Cambridge University Press
65. D.C. Montgomery, *Introduction to Statistical Quality Control* (J. Wiley & Sons, Hoboken, NJ, 1996) ISBN: 0-471-30353-4
66. Naphcare, What skills are important for a clinician in this type of work environment? (2016)
67. G. Nordin, R. Dybkaer, U. Forsum, X. Fuentes-Arderiu, F. Pontet, Vocabulary on nominal property, examination, and related concepts for clinical laboratory sciences (IFCC-IUPAC recommendations 2017). Pure Appl. Chem. **90**(5), 913–935 (2018)
68. L.R. Pendrill, Operating 'cost' characteristics in sampling by variable and attribute. Accred. Qual. Assur., **13**, 619–631, (2008). https://doi.org/10.1007/s00769-008-0438-y
69. L.R. Pendrill, Risk assessment and decision-making risk assessment and decision-making, in *Theory and Methods of Measurements with Persons*, ed. by B. Berglund (Stockholm, SE),

G. B. Rossi (Genoa, IT), J. Townsend (Bloomington, IN), L. R. Pendrill (Borås, SE), (Psychology Press, Taylor & Francis, 2010) ISBN: 978-1-84872-939-1

70. L.R. Pendrill, R. Emardson, B. Berglund, et al., Measurement with persons. A Eur Network **5**, 42–55 (2010)
71. L.R. Pendrill, *El ser humano como instrument de medida* (e-medida, 2014a)
72. L.R. Pendrill, Man as a measurement instrument. NCSLI Measure J. Meas. Sci. **9**, 24–35 (2014b)
73. L.R. Pendrill, Using measurement uncertainty in decision-making & conformity assessment. Metrologia **51**, S206 (2014c)
74. L.R. Pendrill, Assuring measurement quality in person-centered healthcare. Meas. Sci. Technol. **29**(034003) (2018)
75. L.R. Pendrill, Quality Assured Measurement – Unification across Social and Physical Sciences, in *Springer Series in Measurement Science and Technology*, (2019) ISBN: 978-3-030-28695-8 (e-book)
76. L.R. Pendrill, S. Cano, T. Köbe, J. Melin, A. Fillmer, *Restitution of Ability and Difficulty from Decision-Making: Metrology of Human-Based Perceptions* (Measurement at the Crossroads: History, Philosophy, and Sociology of Measurement, Paris, 2018)
77. L.R. Pendrill, J. Melin, S. Cano, the NeuroMET Consortium, Metrological references for health care based on entropy, in *19th International Congress of Metrology, Paris (FR)*, (EDP Science: Web of Conference Open Access, 2019) (Sept 2019)
78. L.R. Pendrill, J. Melin, S.J. Cano, *Entropy-Based Explanations of Multidimensionality in Ordinal Responses* (MSMM, 2021)
79. A. Pentland, *Social Physics: How Good Ideas Spread – The Lessons from a New Science* (Penguin, 2014)
80. K. Pesudovs, Item banking: A generational change in patient-reported outcome measurement. Optom. Vis. Sci. **87**(4), 285–293 (2010)
81. W.W. Peterson, T.G. Birdsall, The Theory of Signal Detectability, Part 1. The general theory, *Project* M970, US Signals Corps project no. 29-194B-0 (1953)
82. A. Possolo, Measurement, in *Advanced Mathematical and Computational Tools in Metrology and Testing: AMCTM XI*, ed. by A. B. Forbes et al., (World Scientific Publishing Company, Singapore, 2018), pp. 273–285
83. G. Rasch, On general laws and the meaning of measurement in psychology, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, (University of California Press, Berkeley, 1961), pp. 321–334
84. K.E. Roach, Measurement of health outcomes: Reliability, validity and responsiveness. JPO **18**, p8 (2006)
85. G.B. Rossi, *Measurement and Probability – A Probabilistic Theory of Measurement with Applications* (Springer Series in Measurement Science and Technology – Springer, Dordrecht, 2014). https://doi.org/10.1007/978-94-017-8825-0
86. T. Salzberger, M. Koller, The direction of the response scale matters – Accounting for the unit of measurement. Eur. J. Mark. **53**, 871–891 (2019). https://doi.org/10.1108/EJM-08-2017-0539
87. C.J. Soto, O.P. John, S.D. Gosling, J. Potter, The developmental psychometrics of big five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. J. Pers. Soc. Psychol. **94**, 714–737 (2008)
88. C. Spearman, The proof and measurement of association between two things. Am. J. Psychol. **15**, 72–101 (1904)
89. C. Spearman, Correlation calculated from faulty data. Br. J. Psychol. **3**, 271–295 (1910)
90. M.H. Stone, Quality control in testing. Popular Meas **4**(1), 15–23 (2002)
91. J.A. Tukey, Chapter 8, Data analysis and behavioural science, in *The collected works of John A Tukey, Volume III, Philosophy and principles of data analysis: 1949–1964*, ed. by L. V. Jones, (University of North Carolina, Chapel Hill, 1986)

92. V. Turetsky, E. Bashkansky, Testing and evaluating one-dimensional latent ability. Measurement **78**, 348–357 (2015)
93. J. Uher, Quantitative data from rating scales: An epistemological and methodological enquiry. Front. Psychol. **9**, 2599 (2018)
94. M.K. Walton, J.H. Powers III, J. Hobart, D. Patrick, P. Marquis, S. Vamvakas, M. Isaac, E. Molsen, S. Cano, L.B. Burke, Clinical outcome assessments: Conceptual foundation – Report of the ISPOR clinical outcomes assessment – Emerging good practices for outcomes research task force. Value Health **18**, 741–752 (2015)
95. D.S. Weitzner, M. Calamia, Serial position effects on list learning tasks in mild cognitive impairment and Alzheimer’s disease. Neuropsychology **34**, 467–478 (2020)
96. WINSTEPS® Manual, (accessed September 21 2022). https://www.winsteps.com/a/Winsteps-Manual.pdf
97. B.D. Wright, Comparing factor analysis and Rasch measurement. Rasch Meas Trans **8**(1) (1994)
98. X. Wu, J. Li, N. Ayutyanont, H. Protas, W. Jagust, A. Fleisher, E. Reiman, L. Yao, K. Chen, for the Alzheimer’s Disease Neuroimaging Initiative, The receiver operational characteristic for binary classification with multiple indices and its application to the neuroimaging study of Alzheimer’s disease. IEEE/ACM Trans. Comput. Biol. Bioinform. **10**, 173–180 (2013)
99. C. Zingg, G. Casiraghi, G. Vaccario, F. Schweitzer, What is the entropy of a social organization? Entropy **21**(9), 901 (2019)
100. J. Zubin (ed.), *Experimental Abnormal Psychology* (Columbia Store, New York, 1955), pp. 2–28. (Mimeo)

Chapter 12
Measurement Systems, Brilliant Processes, and Exceptional Results in Healthcare: Untapped Potentials of Person-Centered Outcome Metrology for Cultivating Trust

William P. Fisher Jr.

Abstract An historic shift in focus on the quality and person-centeredness of health care has occurred in the last two decades. Accounts of results produced from reinvigorated attention to the measurement, management, and improvement of the outcomes of health care show that much has been learned, and much remains to be done. This article proposes that causes of the failure to replicate in health care the benefits of "lean" methods lie in persistent inattention to measurement fundamentals. These fundamentals must extend beyond mathematical and technical issues to the social, economic, and political processes involved in constituting trustworthy performance measurement systems. Successful "lean" implementations will follow only when duly diligent investments in these fundamentals are undertaken. Absent those investments, average people will not be able to leverage brilliant processes to produce exceptional outcomes, and we will remain stuck with broken processes in which even brilliant people can produce only flawed results. The methodological shift in policy and practice prescribed by the authors of the chapters in this book moves away from prioritizing the objectivity of data in centrally planned and executed statistical modeling, and toward scientific models that prioritize the objectivity of substantive and invariant unit quantities. The chapters in this book describe scientific modeling's bottom-up, emergent and evolving standards for mass customized comparability. Though the technical aspects of the scientific modeling perspective are well established in health care outcomes measurement, operationalization of the social, economic, and political aspects required for creating new degrees of trust

W. P. Fisher Jr. (✉)
BEAR Center, Graduate School of Education, University of California, Berkeley, CA, USA

Research Institutes of Sweden, Gothenburg, Sweden

Living Capital Metrics LLC, Sausalito, CA, USA
e-mail: wpfisherjr@livingcapitalmetrics.com

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3_12

in health care institutions remains at a nascent stage of development. Potentials for extending everyday thinking in new directions offer hope for achieving previously unattained levels of efficacy in health care improvement efforts.

Keywords History · Philosophy · Quality improvement · Participatory social ecologies

12.1 Introduction: The Role of Measurement in the Shift to Quality in Health Care

Over 20 years ago, the Institute of Medicine (IOM), now known as the National Academy of Medicine (NAM), released two landmark reports, *To Err is Human* [171] and *Crossing the Quality Chasm* [172]. Berwick and Cassel [26] reflect on the legacy of these reports. They trace out some of the developments that have followed upon the evidence presented as to the seriousness of problems associated with quality in health care in the United States. Even though Berwick and Cassel do not include measurement theory and practice in their account, their observations nonetheless speak directly to the problems addressed by improved quantitative methods. The upshot is that persistent quality improvement challenges in health care may be transformed into entrepreneurial opportunities in the context of person-centered outcome metrology.

While noting that progress in some areas has been significant, Berwick and Cassel make several observations that set the stage for crystallizing a new framework for quality improvement, one providing a needed transformation of the system redesigns originally provoked by the IOM/NAM reports. The relevant lessons learned over the last 20 years, according to Berwick and Cassel, include:

- Wholesale, systemic improvement in quality of care is difficult to bring to scale.
- Improvements tend to be local and not generalized.
- Quality improvement increasingly takes a back seat to financial pressures.
- A driving economic precept of other industries – namely, controlling costs by improving quality – seems unworkable in health care.
- Value-based pay-for-performance schemes have proliferated but have scarcely dented the dominant fee-for-service model.
- Accountability for outcomes has perversely led to serious, negative impacts on clinician morale, and has not led to the desired progress in quality and safety.
- A balance between the critical need for accountability, on the one hand, and supports for a culture of trust focused on growth and learning, on the other, seems out of reach.
- New payment models that would set more constructive priorities have proven politically unpalatable in light of the way cost controls seem to automatically implicate some form of rationing.

One theme in particular repeats itself in a subtle and unnoticed way throughout Berwick and Cassel's account. That theme concerns the unquestioned assumption that quality improvement, accountability, system redesigns, health care economics, and measurement are all inherently and definitively approached via centrally planned data analytics and policy implementations. Conceptual tools and methods are always applied from the outside in and the top down. No other alternative is ever mentioned or considered.

This unexamined assumption is nearly all-pervasive within the paradigm informing quality improvement efforts in person-centered outcome measurement [282, 300]. Questions have been raised as to the shortcomings of this paradigm, and alternatives often based in Actor-Network Theory (ANT) and related perspectives have been proposed [23, 40, 70, 85, 185, 252, 264, 276]. Though measurement and metrological infrastructures are of central importance in Actor-Network Theory [33, 188, 259, 340], only a few accounts referencing it [101–103, 114, 128] focus on the opportunities and problems addressed in advanced measurement theory and practice. Conversely, efforts aimed at establishing metrological systems for person-centered outcome measurements rarely recognize or leverage relevant concepts from Actor-Network Theory [17, 18, 51, 92, 273], though they are key to understanding the propagation of representations across instruments across boundaries. Might opportunities for producing practical results multiply rapidly if advanced measurement and metrology were blended with the ANT-oriented approaches to lean thinking and quality improvement that have been recently initiated in health care [85, 264]?

Actor-Network Theory and related perspectives in science and technology studies show that, in science, new understandings emerge progressively as data patterns cohere into measured constructs, are explained by predictive theories, are then packaged in instruments calibrated and traceable to unit standards, and then distributed to end users. Networks, or, better, multilevel complex ecosystems, of persons equipped with such technologies are enabled to coordinate their decisions and behaviors without having to communicate or negotiate the details of what they see. This is the effect referred to in economics as the "invisible hand;" it becomes possible only in the wake of efforts focused on identifying persistently repeatable patterns in the world that can then be represented in standardized symbol systems. This virtual harmonization of decisions and behaviors is the purpose of measurement.

This chapter proceeds by providing some background on the differences between quantitative methods defined primarily in statistical terms and those defined in relation to metrological potentials. As will be shown, these differences are paradigmatically opposed in their fundamental orientations and in the kinds of information they produce. After introducing the metrological paradigm via this contrast with existing practice, some aspects of its potential for revitalizing the measurement and management of quality improvement efforts in health care are elaborated.

12.2 Modern Statistical Approaches vs. Unmodern Metrological Approaches

As the early psychometrician, Thurstone [321, p. 10] put it in 1937, mathematics is not just an analytic tool, it is the language in which we think. Thurstone sought to make mathematics the language of measurement and psychological science not just in the sense of expressing analytic results but as substantiating the meaning of the relationships represented in numbers. Broadly put, Thurstone was trying to counter the modern Western worldview's Cartesian and positivist subject-object dualism. Here, subjective perceptions are pitted against an objective world assumed to exist independent of any and all human considerations. This is the worldview informing the assumption that mathematics provides analytic tools for application to problems. Understanding how mathematics constitutes the language in which we think entails a quite different perspective.

Alternatives to the modern worldview include postmodern and unmodern perspectives [189, 191, 192, 196]. Postmodern, deconstructive logic notes the historical evolution and theory-laden dependencies of the modern worldview's presentation of transcendent universals [198, 199]. The modern and postmodern quickly become locked in futile and unproductive conflict, however, as their implicit assumptions set up irreconcilable differences. As Latour [191, p. 17] says, "postmodernism is a disappointed form of modernism" that still proceeds with a certain logic but abandons hope of ever arriving at useful generalizations. The modern and the postmodern each offer their distinctive though opposed contributions: laws and theories are indeed powerfully predictive at the same time they are unrealistic formalisms detached from local situations; what counts as data changes as theory changes; and of course, theories are tested by but never fully determined by data. The facts of these empirically observable associations are of no use, however, in resolving the debate.

But the situation changes when the role of standards in language and the history of science are included in the account. Even the philosopher most closely associated with postmodern deconstructions, Derrida, understood this, saying, "When I take liberties, it's always by measuring the distance from the standards I know" [81, p. 62]. He had previously recognized the need to be able, like Levi Strauss in his anthropology, "to preserve as an instrument something whose truth value he [Levi-Strauss] criticizes" [80, p. 284], such as the standards of language. And just as postmodernism has to accept a role for linguistic standards in deconstruction, so also must modernist forms of strong objectivity accept that, though Ohm's law may well be universal, it cannot be proven without a power source, cable, and an ohmmeter, wattmeter, and ammeter [188, p. 250].

Even more fundamentally, given that language itself is a human construction associating arbitrary sounds and shapes with ideas and things, if even linguistic technology was taken away, one would have no means at all of formulating ideas on, thinking about or communicating the supposedly self-evident and objectively independent natural world. The fact that nothing is lost in integrating the modern metaphysics of a universally transcendent nature with postmodern relativism in an unmodern semiotics is recognized by Ricoeur [294]. He similarly points at the objectivity of text as a basis for explanatory theory not derived from a sphere of events assumed to be natural (existing completely independent of all human conception or interests), but compatible with that sphere. A semiotic science taking language as a model then does not require any shift from a sphere of natural facts to a sphere of signs: "it is within the same sphere of signs that the process of objectification takes place and gives rise to explanatory procedures" [294, p. 210].

Latour also notes that the unmodern perspective loses nothing in its pragmatic idealism that was claimed by modernism, while also offering a new path forward for research. As he says [192, p. 119],

> To speak in popular terms about a subject that has been dealt with largely in learned discourse, we might compare scientific facts to frozen fish: the cold chain that keeps them fresh must not be interrupted, however briefly. The universal in networks produces the same effects as the absolute universal, but it no longer has the same fantastic causes. It is possible to verify gravitation 'everywhere', but at the price of the relative extension of the networks for measuring and interpreting.

It is also refreshing to have a frank acknowledgment of how difficult and expensive it is to bring new things into language and keep them there [142, 188, p. 251; 247, 248, 278], as opposed to the way modernist metaphysics renders invisible the processes by which the facts of nature are made to seem obviously, "naturally," freely, and spontaneously self-evident and available [107, 299].

Dewey [83, 84] and Whitehead [331, 332] anticipate important aspects of the unmodern alternative articulated largely by Latour and others in the domain of science and technology studies and Actor-Network Theory. The unmodern perspective brings the instrument to bear in a semiotics of theories, instruments, and data based in the semantic triangle of ideas, words, and things. This extension does nothing but accept the historical development of written language as a model of what Weitzel [329] referred to as "a perfect standardization process" in his study of the economics of standards in information networks. As most readers are likely unfamiliar with both unmodern ideas and semiotics, it will be worthwhile to linger a bit on this topic, and provide some background.

In the history of science, research brings new things into language via what Wise [103, 340] describes as a two-stage process. New constructs initially act as agents compelling agreement among observers as to their independent existence as something repeatedly reproducible. In the second stage, the agent of agreement is transformed into a product of agreement made recognizable and communicable in the standardized terms of consensus processes. These processes culminate in metrological networks of instruments calibrated to fit-for-purpose tolerances.

Our modern Western cultural point of view usually conceives of society as making use of symbols and technologies, but closer examinations of the historical processes of cultural change [190, 191, 193–195, 330] show that symbolization exerts more influence over society than vice versa [20, 48, 176]. Features of the world, after all, do not become socially significant and collectively actionable until they are symbolized in shared representations. Communications and metrological networks are, then, specialized instances of the broader political economy of societies structured by means of shared symbol systems: “social reality is fundamentally symbolic” [294, p. 219]. In Alder’s [5] terms, “Measures are more than a creation of society, they *create* society.”

Because we are born into a world of pre-existing languages, an economy of thought [16, 133, 134] facilitates communication by lifting the burden of initiation [137, p. 104] and absorbing interlocutors into a flowing play of signifiers. That is, language and its extensions into science via metrology are labor-saving devices in the sense that they relieve us of the needs to create our own symbols systems, and to then translate between them. In this sense, as Gadamer [137, p. 463] says, it is truer to say that language speaks us than to say that we speak language.

Wittgenstein [341, p. 107] concurs, adding, “When I think in language, there aren’t ‘meanings’ going through my mind in addition to the verbal expressions: the language is itself the vehicle of thought.” Attending to the unity of thing and thought in language constitutes the Hegelian sense of authentic method as *meta-odos*: how thought follows along after (*meta*) things on the paths (*odos*) they take of their own accord [137, pp. 459–461; 143, 160, p. 63; 256–258]. Understood in its truth, method embodies a performative logic of things as they unfold in the relational back and forth of dialogue.

As has been established at least since the work of Kuhn [182, 183] and Toulmin [323, 324], methods always entail presuppositions that cannot be explicitly formulated and tested. The sense of method that focuses on following rules must necessarily always fall short in its efforts at explanatory power and transparency [137]. Going with the flow of the mutual implication of subject and object, however, enables a tracking or tracing of the inner development of the topic of the dialogue as it proceeds at a collectively projected higher order complexity. Persistently repeated patterns in question-and-answer dialogues facilitated by tests, assessments, surveys, etc. are the focus of measurement investigations [104–106, 110, 120].

An example of how measurement scaling methodically anticipates, models, investigates, maps, documents, and represents the movement of things themselves experienced in thought is given in Fisher’s [106] interpretation of Wright and Stone’s study of the Knox Cube Test [317, 356]. This and related tests, such as the Corsi Block Test, have been taken up as candidates for metrological standardization in recent efforts in Europe [50, 229, 270, 286, 287]. This work may play an important role in creating and disseminating on broad scales a powerful new class of phenomenologically rich methods and instruments.

In the same way that we think and communicate only in signs [38, 39, 61, 71, 79, p. 50; 83, p. 210; 268, 269, p. 30; 294, pp. 210, 219; 332, p. 107; 341, p. 107], so, also, is measurement the medium through which mathematics functions as the language of science [255]. Instruments are the media giving expression to science’s mathematical language. In Latour’s [188, pp. 249–251] words,

> Every time you hear about a successful application of a science, look for the progressive extension of a network.... The predictable character of technoscience is entirely dependent on its ability to spread networks further. ...when everything works according to plan it means that you do not move an inch out of well-kept and carefully sealed networks. ... Metrology is the name of this gigantic enterprise to make of the outside a world inside which

> facts and machines can survive. … Scientists build their enlightened networks by giving the outside the same paper form as that of their instruments inside. [They can thereby] travel very far without ever leaving home.

Dewey [83] more broadly described the situation, saying:

> Our Babel is not one of tongues but of the signs and symbols without which shared experience is impossible. … A fact of community life which is not spread abroad so as to be a common possession is a contradiction in terms. ...the genuine problem is that of adjusting groups and individuals to one another. Capacities are limited by the objects and tools at hand. They are still more dependent upon the prevailing habits of attention and interest which are set by tradition and institutional customs. 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.
>
> A more intelligent state of social affairs, one more informed with knowledge, more directed by intelligence, would not improve original endowments one whit, but it would raise the level upon which the intelligence of all operates. The height of this level is much more important for judgment of public concerns than are differences in intelligence quotients. As Santayana has said: 'Could a better system prevail in our lives a better order would establish itself in our thinking.'

Taking up the same theme, Whitehead [332] noted that the then-recent changes in the science of physics came about because of new instruments that transformed the imaginations and thoughts of those trained in using them. Scientists had not somehow become individually more imaginative. He [331, p. 61] previously remarked on the way civilization advances by distributing access to operations that can be successfully executed by persons ignorant of the principles and methods involved. Much the same point is made in developmental psychology's focus on the fact that "cultural progress is the result of developmental level of support" [63; 327].

We stand poised to transform the developmental levels of support built into the cognitive scaffolding embedded in today's social environments. Our prospects for success depend extensively on our capacities for building trust, something that is quite alien to numbers situated in isolated contexts disconnected from substantive diagnostic, theoretical, or meaningful values [262, 280, p. 144].

Building trust by connecting numbers with widely distributed, quality assured, reliable, dependable, repeatable, and meaningful structural invariances will inevitably make use of Rasch's probabilistic models for measurement, which are eminently suited to satisfying the need for metrological networks of assessment- and survey-based instruments [102, 108, 109, 111, 112, 214, 216, 271, 272]. Wright [352], a primary advocate and developer of Rasch's mathematics, noted not only that "science is impossible without an evolving network of stable measures," he also made fundamental contributions to theory; historical accounts; modeling; estimation; reliability, precision, and data quality coefficients; instrument equating; software; item bank development and adaptive administration; report formatting; the creation of new professional societies and journals; and applications across multiple fields [116, 337]. Over the last 40 years and more, research and practice in education, psychology, the social and environmental sciences, and health care have benefited greatly from the advances in measurement developed by Wright, his students, and colleagues [4, 8, 9, 12, 13, 29, 89, 90, 126, 200, 210, 225, 295, 335].

Though much has been accomplished in the way of consolidating theoretical conceptions and descriptions of how sociocognitive infrastructures are embedded in cultural environments [32, 35, 124, 177], practical applications of these ideas have scarcely begun. Focusing on the cultivation of trust and dependable measurement standards may go a long way toward starting to reverse many of the communications problems in the world today. Scientifically defensible theories of measured constructs create an important kind of trust as explanatory models persistently predict instrument performance. Experimental results of instrument calibrations across samples over time and space similarly support the verifiable trust needed for legal contracts and financial accounting. Finally, distributed networks of trust may come to be embodied in end users who experience the reliable repetition of consistent results, who are able to see themselves and others in their data, and the data in themselves and others.

12.3 Creating Contexts in Health Care for Success in Lean Thinking

With that background in mind, how might an unmodern, metrological point of view lead to transformed, trustworthy quality improvement systems in health care? Answers are suggested with particular pointedness by repeated conceptual disconnections between quality, costs, and reduced waste in health care's lean applications.

"Lean thinking" [279, 343] focuses on systematically removing processes that do not contribute to outcome quality. Lean processes are those from which ineffectual and costly, and so wasteful, factors have been eliminated, often via experimental comparisons. Manufacturing industries' quality improvement and cost management successes with lean thinking have long been of interest in health care [27, 69]. The Toyota Production System is widely taken as a preeminent example of success in lean thinking. That success led Fujio Cho, a past Chairman of Toyota Motors, to remark, "We get brilliant results from average people managing brilliant processes. Our competitors get average results from brilliant people working around broken processes" (as quoted in [301], p. 84). Implementations of lean thinking over the years in health care, however, have tended to be piecemeal, isolated within departments, with little in the way of interdisciplinary boundary crossing, as was shown in a recent literature review [3]. As is also the case in education [87], because of these problems, health care organizations have a propensity to learn the same lessons over and over again, as a kind of "organizational amnesia" [261].

The goals for research capable of crossing boundaries include supplying evidence of more coherent knowledge, and taking a more integrated, authentic view of life [236]. A productive way of framing these goals pairs low and high levels of conceptual learning ("know why") with low and high levels of operational learning ("know how") to categorize types of situations in which organizations are able and unable to learn across boundaries [187]. Though this scheme has been productively applied in surgical contexts [132, 275], health care generally seems to be vacillating between firefighting, where conceptual and operational learning are both low; unvalidated theories, where conceptual learning is high and operational is low; and artisan skills, where conceptual learning is low and operational is high.

Berwick and Cassel's reflections on quality improvement, and Akmal and colleagues' [3] account of lean thinking, document the nearly complete failure in health care to achieve operationally validated theories, where both "know how" and "know why" are maximized. In a relevant insight, the blindered perspective on quality in health care is attributed in part to a "tool-myopic" thinking that undermines the generalizability of lean processes [3]. Going past Akmal et al.'s insight, we can see that this near sightedness comes from viewing measured outcomes only through the lenses of ordinal scores. When the meaning of numbers is tied to specific items, measurement cannot fulfill its role as the external scaffolding needed to facilitate socially distributed remembering and embedded cognitive supports, to adopt the language of Sutton and colleagues [318].

This myopia affecting quality improvement in health care is, then, ostensibly caused by the design limitations of quality-of-care instruments and measurement systems [130, 161, 356]. The theory of action following from ordinal metrics cannot tell people what to do in general; they can only inform the specific operations included in the score. As the economist, Joseph Stiglitz, put it, "What you measure affects what you do. If you don't measure the right thing, you don't do the right thing" (quoted by Goodman [146] in the *New York Times*; also see [131]). Stiglitz was speaking to the concrete content of what is measured, but including the right content is insufficient to the task of measurement. The content must be organized and configured in particular ways if it is to be meaningfully measured and appropriately actionable. Stiglitz' observation ought then to be stated as, "The *way* you measure affects what you do. If you don't measure the right *way*, you don't do the right thing."

What this means can be seen in the fallacy of misplaced concreteness [332, pp. 52–58] taken for granted as an unavoidable complication in most survey and assessment applications in health care. Though the intention is to obtain information applicable in general across departmental, organizational, and other boundaries, restricting the conceptualization and operationalization of measurement to numeric counts focuses attention counterproductively on the specific things counted. The maxim that we manage what we measure becomes in this context unintentionally ironic, as managers wind up focusing on bean counting instead of attending to their mission. Managing the numbers becomes a consequence evident in Berwick and Cassel's list of lessons learned from quality improvement efforts in health care over the last 20 years.

Mainstream measurement methods in health care assume that assigning scores, counting events, summing ratings, or computing percentages suffice as measurement. Online searches of the contents of three recent textbooks on quantitative methods in health care [42, 263, 267], for instance, show that no mention is made of key words (such as invariance, statistical sufficiency, metrology, traceability; consensus standards, Rasch models, etc.) associated with measured quantities, as opposed to numeric statistics. As is the usual situation throughout psychology and the social sciences, quantitative methods in health care, more often than not, merely assume numbers to represent unit amounts of something real in the world, and never formulate or test the hypothesis that such quantities exist [10, 233, 314, 349, 352]. For instance, a January 2022 search of Google Scholar shows that only 2 of 74 peer-reviewed articles citing Nolan and Berwick [253] on "all-or-none measurement" even mention invariance or test the quantitative hypothesis [82, 175].

Concrete models of outcome quality implicated when measurement is misconceived as merely numeric are not the only option, however. Patterns of learning, healing, development, and growth demonstrated as invariantly structured across samples and instruments have been available and in use in health care for decades. A Google Scholar search of "Rasch model" and "health outcomes" in January of 2022 gave over 3000 results. A selection of even early examples shows a wide range of applications across diagnostic, clinical, treatment, and licensure/certification contexts [22, 24, 29, 30, 49, 54, 68, 97, 154, 163, 178, 209, 304, 354].

Most applications of Rasch's models for measurement continue to operationalize them statistically as centrally planned exercises in data gathering and analysis. Fit to demanding measurement models may be highly prized in some quarters, but confidence in generalizing the results seems to be lacking. This remains the case despite instances in which the invariance of constructs has been demonstrated repeatedly over samples, over periods of years involving multiple instruments, or across multiple applications of the same instrument [52, 100, 102, 127, 148, 296, 305, 328, 334]. Exceptions to this pattern focus on developing and deploying adaptively administered item and scale banks [19, 285, 319, 326, 353], and self-scoring forms interpretable in measurement terms at the point of use [28, 34, 72, 165, 179, 201, 204, 274, 325, 342, 344, 351].

Another exception to the usual data-focused approach to measurement is the Low Vision Research Network [41, 53], which is based in an extended instrument equating and item bank development program [91, 145, 217–224]. Similar kinds of networks termed STEM (science, technology, engineering, and mathematics) ecosystems coordinating education and workforce development needs have mathematically modeled stakeholder empowerment to integrate measurement and management dialogues among formal and informal educators, technology companies, city and state governments, funding agencies, and philanthropists [237–240]. Finally, the European NeuroMet consortium [50, 229, 230, 270, 286, 287] is developing metrological standards for rehabilitative treatments of neurodegenerative conditions affecting cognitive performance. The instruments being studied in this project include variations on the Knox Cube Test, which was taken as a primary example of a simple exercise in scaling and theory development by Wright and Stone [317, 356].

This latter domain of attention span and short-term memory is of additional interest for its intensive focus on theory development [230, 311]. Rasch's models facilitate not only *empirical* experiments testing whether concrete numeric counts falsify hypotheses of abstract measured quantities, they also structure and inform *theoretical* experiments testing whether the abstract scale is explained at a higher-order formal complexity affording explanatory and predictive power [77, 89, 94, 284, 311–315, 334, 335].

Automatic item generation [15, 88, 144, 166, 181, 277, 307, 313–315] becomes a possibility when theory advances to the point that empirically estimated item locations correspond usefully with theoretical predictions. The high cost of composing new test and assessment items is dramatically reduced by automation [181, 313]. The economy of thought structured by the model of language attains otherwise inaccessible degrees of market efficiency when single-use items formulated on the fly by a computer produce the response frequencies predicted by theory. Routine reproduction of these results in multiple domains will be key to altering the prevailing habits of attention and the overall social level of intelligence in the manner described by Dewey [83]. Instances of the stable reproduction of scales supported by theory and evidence, such as that described by Williamson [334], will be key to building out verified and justified investments of trust in numbers.

Measurement methodologies in use today in health care range, then, from ordinal item-dependent scores to interval analysis-dependent scales, and from there to metrologically traceable quantities. The larger modern cultural context accepts the objectivity of data as a record of events assumed to suffice as criterion for reproducibility. It seems that neither repeated conceptual critiques of these assumptions [59, 152, 153, 186, 228, 233, 302, 303, 349], nor readily available operational alternatives [7, 12, 95, 207, 246, 288, 347, 352] have had much impact on most researchers' methodological choices. Is it possible to imagine any kind of a compelling motivation that would inspire wide adoption of more meaningful, rigorous, and useful measurement methods? The logic and examples developed and produced over the last 60 years have failed to shift the paradigm in the way that might have been expected, as has been recognized for some decades [58, 246]. Might there be other ways of making the superiority and advantages of some forms of information so attractive that few would miss the opportunity to use them?

12.4 Tightening the Focus on Metrological Potentials in Health Care

A plausible reason why conceptually and operationally superior measurement methods have not been more widely adopted is that the existing systems of incentives and rewards favor the status quo. It is often said that the inequities and biases of today's institutions are systemic, that the institutions themselves seem designed to disempower individuals not born into advantageous circumstances [6, 245]. Mainstream methods of measurement and management prioritize concrete responses to the particular questions asked in local circumstances, doing so in ways that systematically undercut possibilities for creating the collective forms of information needed for reconfiguring our institutions. It may be that we will be unable to transform the systems of incentives and rewards dominating those institutions until we cultivate new ecologies of cross-disciplinary relationships that displace the existing ones by offering more satisfying substantive and economic outcomes. The root question is how to extend the economy of thought facilitated by language into new sciences and associated economic relationships.

That is, a new science of transdisciplinary research and practice is needed if we are to succeed in addressing the complex problems we face. The complexity of articulating how collective forms of knowledge are measured and managed, however, may cause even those focused squarely on the problem to nonetheless revert to established habits of mind and conceptual frameworks. The end result is often, then, a mere repetition of modern metaphysics in yet another take on concrete, individual-level solutions, as has been the case in some recent efforts [44, 316].

A number of fields, however, have taken steps toward practical understandings of collective complexities, with a particularly compelling convergence of perspectives:

- Science and technology studies offer the concept of the boundary object [37, 308, 310] and the trading zone [140], in which meanings are recognized as simultaneously concrete and abstract, and as co-produced in tandem with standardized technologies ranging from alphabets and phonemes to clocks, thermometers, and computers [155, 174, 189].
- Developmental psychology, similarly, offers a theory of hierarchical complexity [60, 64, 73, 96] that shows how thinking is general and specific at the same time; developmental transitions across 15 levels of complexity spanning the entire range of variation in the human life cycle are well documented.
- Cognitive psychology also has in recent years come to be concerned with the ways in which thinking does not occur solely in the brain, neurologically, but depends materially on the sophistication of external supports like alphabets, phonemes, grammars, and other technologies embedded in the external environment [168, 169, 318].
- Studies in the history of science similarly expand on the role of models as material artifacts mediating innovative advances [249–251]. Here again, formal conceptual ideals are represented in arbitrary standardized abstractions useful in negotiating local meanings in unique concrete circumstances.
- Semiotics operationalizes the pragmatic consequences of knowing that all thought takes place in signs, with language as its vehicle, or medium [38, 39, 71, 260, 265, 266, 268, 269, 298]. The semiotic triangle's thing-word-idea assemblages integrate concrete, abstract, and formal levels of complexity in practical, portable ways readily extended [115, 117] into the data-instrument-theory combinations documented in science and technology studies [1, 157, 170].

The overarching shared theme evident across these fields points at the need for distributed networks of instruments capable of operationalizing – simultaneously – formal explanatory power, abstract communications systems, and concrete local meaningfulness [115, 117].

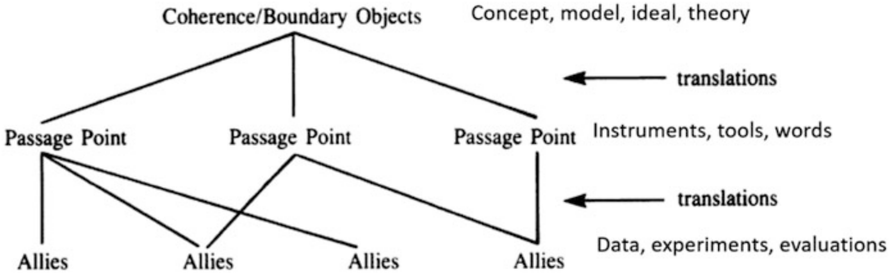

Coherence/Boundary Objects
Concept, model, ideal, theory
translations
Passage Point
Passage Point
Passage Point
Instruments, tools, words
translations
Allies
Allies
Allies
Allies
Data, experiments, evaluations

Fig. 12.1 Levels of complexity in knowledge infrastructures (Ref. [309, p. 390]; annotated)

Figure 12.1, adapted from Star and Griesemer [309], schematically illustrates the concrete data-level alliances of participants in knowledge infrastructures, their relations to abstract consensus standards' obligatory passage points, and the encompassing umbrella of formal, theoretical boundary objects. Measurement research overtly addresses each of these levels of complexity. Systems integrating all three levels may become themselves integrated in meta-systematic systems of systems, or in paradigmatic supersystems [60, 115]. Associating substantive and reproducible patterns of variation consistently revealed in measurement research with consensus standards, and building these standards into applications across fields, may provide the direction needed for designing knowledge infrastructures capable of compelling adoption of improved methods in health care quality improvement. That is, translation networks' alliances, obligatory passage points, and boundary objects define contexts in which systems elevate the overall level of intelligence and imagination, enabling everyday people to produce brilliant results. Lean thinking is, then, the apotheosis of the more broadly defined translation networks actualizing the economy of thought facilitated by language.

The history of economics points at the necessary roles played by measurement across legal, governance, communications, scientific, and financial institutions [2, 5, 14, 21, 25, 78, 254, 283, 320, 333]. Where the transcendental ideals of modern thinking assume that the objective world's real and independent existence suffices to inform the assignment of numbers in measurement, the pragmatic unmodern recognition of the essential, difficult to craft, and highly expensive roles played by instruments and standards suggests a need for new directions.

Legal, financial, technical, and social networks and standards make nature seem given and obvious by weaving, as Callon [47] puts it, a "socionature:" an environment in which social associations mediated by mutually supportive shared symbol systems make things appear to be part of an objective reality independent of human concerns. Reality actually is, of course, independent of the concerns and interests of individual people, but it exists as a shared and accessible reality only when human interests collectively project higher order complexities and coordinate them in systems of trusted interdependent symbolic representations. Measurement research needs to feed into and become embedded in the legal, financial, social, and metrological infrastructures as part of the extension of everyday thinking and language into new domains.

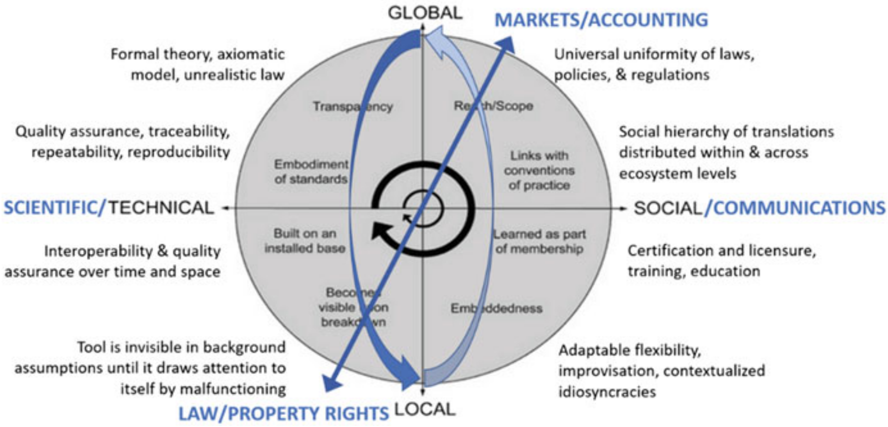

GLOBAL
MARKETS/ACCOUNTING
Formal theory, axiomatic model, unrealistic law
Universal uniformity of laws, policies, & regulations
Transparency
Reach/Scope
Quality assurance, traceability, repeatability, reproducibility
Embodiment of standards
Links with conventions of practice
Social hierarchy of translations distributed within & across ecosystem levels
SCIENTIFIC/TECHNICAL
SOCIAL/COMMUNICATIONS
Built on an installed base
Learned as part of membership
Interoperability & quality assurance over time and space
Certification and licensure, training, education
Becomes visible upon breakdown
Embeddedness
Tool is invisible in background assumptions until it draws attention to itself by malfunctioning
Adaptable flexibility, improvisation, contextualized idiosyncracies
LAW/PROPERTY RIGHTS
LOCAL

Fig. 12.2 Co-produced technical and social infrastructural dimensions. (Adapted from Ref. [36, p. 392])

Figure 12.2 is adapted from Bowker's [36, p. 392] Fig. 29.1; his original version specifies only the vertical Global/Local and horizontal Technical/Social dimensions, and all text apart from those labels is inside one of the encircled four quadrants. The annotations of the quadrants outside of the circle have been added, as also have the expansions of the Technical to include the Scientific, and of the Social to include Communications. Finally, a third, orthogonal, dimension was added to further expand the Social to include Legal and Property Rights, and the financial domain of Markets and Accounting.

Figure 12.3 rotates Fig. 12.2 to reveal the Global and Local dimensions of the Legal/Property Rights and Markets/Accounting domains. The complex resonances of relationships coordinated and aligned in loosely coupled ways across these domains and levels of complexity suggest that different actors playing different roles at different levels are not going to agree entirely on what the object of interest is, or how it functions. Galison [140, 141] found the divergence of opinions about microphysical phenomena across experimentalists, instrument makers, and theoreticians so striking that he felt forced to assert that the disunity of science must play a key role in its success. Similarly, in contrast with the assumption that paradigms define contexts of general agreement in science, Woolley and Fuchs [345, p. 1365] "suggest that a spectrum of convergent, divergent, and reflective modes of thought may instead be a more appropriate indicator of collective intelligence and thus the healthy functioning of a scientific field."

In this context, when diverse alliances implement measurement standards to inform common product definitions and trustworthy expectations of market stability across technical, legal, financial, and communications sectors, it becomes apparent that certain kinds of instruments can make markets [121, 234]. Markets are not, contrary to popular opinion, made merely by exchange activity. Instead, "skilled

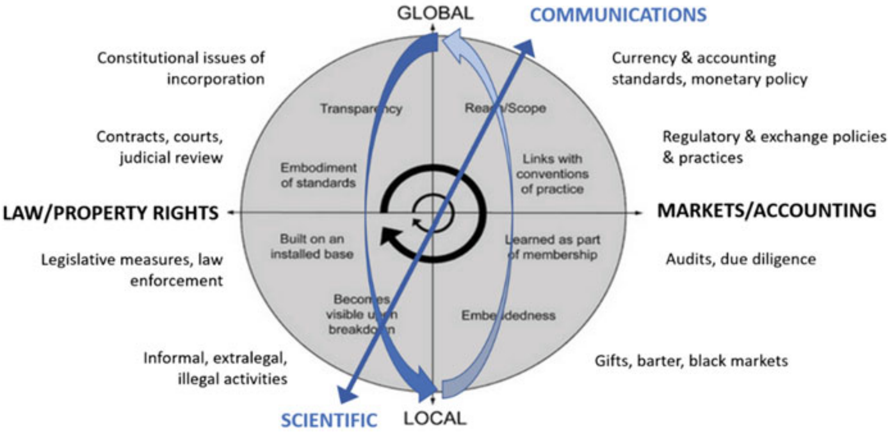

GLOBAL
COMMUNICATIONS
Constitutional issues of incorporation
Currency & accounting standards, monetary policy
Transparency
Reach/Scope
Contracts, courts, judicial review
Regulatory & exchange policies & practices
Embodiment of standards
Links with conventions of practice
LAW/PROPERTY RIGHTS
MARKETS/ACCOUNTING
Built on an installed base
Learned as part of membership
Legislative measures, law enforcement
Audits, due diligence
Becomes visible upon breakdown
Embeddedness
Informal, extralegal, illegal activities
Gifts, barter, black markets
SCIENTIFIC
LOCAL

Fig. 12.3 Co-produced legal and financial infrastructural dimensions. (Adapted from Ref. [36, p. 392])

actors produce institutional arrangements, the rules, roles and relationships that make market exchange possible. The institutions define the market, rather than the reverse" [234, p. 710].

In short, the economy is a metrological project [67, 235]. It is comprised of "a series of competing. . .rival attempts to establish metrological regimes, based upon new technologies of organization, measurement, calculation, and representation" [235, p. 1120]. Because they create a babble of incommensurable numeric comparisons, today's ordinal, scale-dependent technologies quantify health outcomes in ways that organize, measure, calculate, and represent quality improvement opportunities inefficiently and ineffectively. The plethora of incommensurable scales makes it impossible for the instruments to ever disappear into their use as relatively transparent media uniting thing and thought. Instead of drawing attention to themselves only when they malfunction, ordinal scales' lack of theoretical explanations and instrument standards make it impossible for the scales to not foreground themselves as an object of interest.

Because diagnostics and theory are overlooked in the development and use of these metrics [93], and because they are opaque black boxes incapable of providing the needed comparability, they have created a crisis of reproducibility [156, 173], and cannot serve as a basis for trusted relationships.

Advanced measurement, however, opens onto new technologies that appear likely to support broader, more productive, and historically proven approaches to creating competing metrological projects. Though clarified quantitative criteria for scaling interval measurement standards set the stage for improved communication and comparability, the way forward will not be without its challenges. The development of electrical standards, for instance, provides a telling account of the uneven and fraught processes bedeviling the creation of commercially viable metrological traceability for resistance and current metrics [167].

That said, there are pointed motivations for engaging with metrological opportunities in health care that were not of concern in markets focused exclusively on manufactured capital. The need to restore trust in our institutions must certainly be included among the highest priorities in this regard. Human, social, and natural capital markets are currently permeated by systemic biases and inequity, problems made to seem ineradicable by the unexamined assumptions of modern thinking and vested interests in maintaining the status quo. Continuing to assume that objective reality is completely independent of human conceptions (i.e., that the assignment of numbers to observations suffices as measurement, despite the resulting communications chaos) promotes and prolongs the domination of disempowering institutions, making it seem as though trying to alter or transform existing relations of power and authority is a futile quest.

Nothing, however, could be further from the truth than this complacent or hopeless attitude. Where ordinal measurements tied to item-dependent categorical scores cannot provide the comparability needed for creating new institutional arrangements in health care markets, metrologically traceable interval measurements possibly could [112, 117, 119, 121]. But, contrary to modernist assumptions as to individual minds being the seat of creativity, it ultimately will not likely be necessary to persuade and educate individuals as to the need for new methods.

That is, as cross-sector associations informed by advanced measurement results proliferate, it will become increasingly counterproductive to remain disconnected from the networks through which trusted, replicable cost-benefit relationships are collectively organized, measured, calculated, and represented. Where contemporary metrological regimes based in ordinal, concrete representations compete in centrally planned command economies, new regimes integrating concrete, abstract, and formal representations will distribute information and decision-making power to end users throughout coordinated networks and efficient markets. When that happens, it would be wise to bet that empowered individuals – clinicians, educators, managers, patients, etc. – will choose to make use of reliable, quality assured, and comparable precision measurements when making their treatment decisions. In this scenario, researchers and practitioners will ignore new metrological developments at their peril.

12.5 Discussion

Apt warnings as to the hubris of mounting such an effort aimed at improving the human condition are given in Scott's [297] historical account of "high modern" schemes intended to serve the greater good. Scott [297] suggests the integrity of these efforts depends on capacities for:

- enhancing the lives of those who are affected by them;
- being deeply shaped by the values of those participating in them; and

- permitting unique local creative improvisations that may not conform to conceptual ideals.

Even systems designed with the most benevolent of intentions can fail catastrophically in one or more of these three areas, with disastrous results. Scott [297] suggests taking language as a model of ways of effectively achieving these capacities for trustworthy relationships. This effectiveness follows, Scott says, from the way language serves as a medium through which broad principles are continually applied to novel circumstances. It is, as he says, "a structure of meaning and continuity that is never still and ever open to the improvisations of its speakers." Scott is implicitly suggesting that the pragmatic idealism of language ought to serve as a model for how to integrate multiple levels of complexity, where formal idealizations, abstract standards, and concrete local events are in play simultaneously.

Well-designed outcome measurement systems [130, 335, 356] might meet all three of Scott's integrity test criteria to the extent they:

- provide end users with information they need to understand where they are now relative to where they have been, where they want to go, and what comes next in their journey, as in formative assessment [31, 54, 113];
- are calibrated on data, provided by end users themselves, exhibiting patterns of structural invariance informing the interpretation of the measurements [288, 347]; and
- can be managed in their specific idiosyncrasies to take special strengths and weaknesses into account irrespective of the pattern expected by the model [55–57, 130, 147, 164, 197, 202, 203, 227, 357].

These pragmatically ideal terms are also implicated in the correspondences across fields listed above, which also focus, in effect, on taking language as a model. Galison [139, pp. 46, 49], for instance, similarly finds in his ethnographic study of microphysics' communities of theoreticians, instrument makers, and experimentalists that the seeming paradox of locally convergent and globally divergent meanings can be reconciled by seeing that "the resulting pidgin or creole is neither absolutely dependent on nor absolutely independent of global meanings."

Star and Griesemer [309] (also see Star [308], Star and Ruhleder [310]) implicate this linguistic suspension between ideas and things when they explain that:

> Boundary objects are objects which are both plastic enough to adapt to local needs and the constraints of several parties employing them, yet robust enough to maintain a common identity across sites. They are weakly structured in common use, and become strongly structured in individual site use. These objects may be abstract or concrete. They have different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable, a means of translation. The creation and management of boundary objects is a key process in developing and maintaining coherence across intersecting social worlds.

In a study of how universalities are constructed in medical work, Berg and Timmermans [23] came to see that:

> In order for a statistical logistics to enhance precise decision making, it has to incorporate imprecision; in order to be universal, it has to carefully select its locales. The parasite cannot be killed off slowly by gradually increasing the scope of the Order. Rather, an Order can thrive only when it nourishes its parasite--so that it can be nourished by it. . . . Paradoxically, then, the increased stability and reach of this network was not due to more (precise) instructions: the protocol's logistics could thrive only by parasitically drawing upon its own disorder.

Berg and Timmermans provide here an apt description of the model-data relation in the context of probabilistic measurement. The models must be probabilistic because of the need to incorporate imprecision and uncertainty. When collectively projected patterns of invariance are identified and calibrated for use in distributed metrological systems, the locales in which the instruments are used must be carefully selected. Though the structural invariance modeled may eventually be replicated over millions of cases and thousands of items, imprecision and uncertainty are never completely overcome. Ongoing data collection will nourish the articulation of the scale to the extent that new persons and items teach new lessons, and the scale's logistics will continue to thrive and its network will be increasingly stable and extended only to the extent that the stochastic invariance patterns persist.

Probabilistic models of measurement incorporate imprecision and uncertainty as the basis for estimating quantity, and counterintuitively provide a firmer basis for quantification than non-probabilistic models. As was noted by Duncan [86], "It is curious that the stochastic model of Rasch, which might be said to involve weaker assumptions than Guttman uses [in his deterministic models], actually leads to a stronger measurement model." Measurements made via probabilistic modeling are, furthermore, evaluated for their usefulness in terms that provide the precision and information quality needed to support a decision process that takes language as a model by explicitly positing, testing, substantiating, and deploying simultaneously and systematically formal theoretical, abstract standard, and concrete data levels of complexity.

This semiotic explication operationalizes language meta-systematically, reflectively acting on it as an object of intentionally framed decisions, instead of allowing unexamined metaphysical assumptions to shift uncontrollably from prioritizing one or another hidden agenda to another. In Postman's [281] terms,

> If we define ideology as a set of assumptions of which we are barely conscious but which nonetheless directs our efforts to give shape and coherence to the world, then our most powerful ideological instrument is the technology of language itself.

The semiotics of language structure a usually unquestioned metaphysical ideology that, as Burtt [43, p. 229] recognized, is "passed on to others far more readily than your other notions inasmuch as it will be propagated by insinuation rather than by direct argument." Modernist and postmodernist emphases on transcendental uniformity, local relativity, empiricism, operationalism, idealism, instrumentalism, etc. each selectively attend to or ignore one or another of the semiotic domains of language. Taking language altogether as a semiotically integrated model of how things are represented, maintained, and transformed sets up new paradigmatic alternatives [38, 39, 71, 255, 260, 265, 266, 298].

Rasch, though neither a philosopher, a semiotician, an historian of science, nor a developmental psychologist, was well aware that his models for measurement function at formal, abstract, and concrete levels of complexity. He [288, pp. 34–35] recognized his models are, for instance, as unrealistic in their form as the idealizations asserted in Newton’s laws, or in geometry. Just as no one has ever observed the inertial path of an object left entirely to itself unaffected by any forces, so, too, is it impossible to draw or observe the mathematical relationships illustrated in geometric figures. This is readily seen in the fact that a right isosceles triangle with unit sides of 1 has a hypotenuse the length of the square root of 2, an undrawable irrational number. Similarly, circles with a radius of 1 have a circumference of pi, another undrawable irrational number.

Butterfield [45, pp. 16–17, 25–26, 96–98] remarks that no amount of photographically recorded observations made during experiments on actual objects in motion would ever accumulate into the kind of geometric ideal envisaged by Galileo. Instead, he says, that vision requires a different kind of thinking-cap, an imaginative transposition that projects an unrealistic but solvable mathematical relationship that might possibly be made useful. As Burtt [43, p. 39] recognized, what Galileo, Copernicus, Newton, and other early physicists accomplished followed from imagining answers to the question as to “what motions should we attribute to the earth in order to obtain the simplest and most harmonious geometry of the heavens that will accord with the facts?”

Rasch, in effect, imagined that the same question could be usefully posed in the domain of human abilities [118, 122]. It is just as unrealistic to model Pythagorean triangles as it is for measurements of human performance to be functions of nothing but the differences between abilities and the difficulties of the challenges encountered. Rasch [288, pp. 34–35; 292] accordingly emphasized that models are not meant to be true, but must be useful. Rasch geometrizes psychology and the social sciences by conceptualizing the relations of infinite populations of persons exhibiting abilities, functionality, or health in the same frame of reference alongside the universes of all possible challenges to those capacities. He puts on the “different kind of thinking-cap” described by Butterfield and Burtt as beginning from the projection of geometric forms, instead of from the assumption that such forms will result from accumulated observations.

Rasch’s approach stands in stark contrast to the dominant paradigm in health care quality improvement and its unquestioned modernist supposition that empirical observations do in fact accumulate into patterns of lawful behavior. This assumption persists despite the more than six decades that have passed since Kuhn ([184], p. 219; original emphasis, and originally published in 1961) noted that

> *The road from scientific law to scientific measurement can rarely be traveled in the reverse direction.* To discover quantitative regularity one must normally know what regularity one is seeking and one's instruments must be designed accordingly; even then nature may not yield consistent or generalizable results without a struggle.

Rasch’s models make the laws of measurement readily accessible, and offer a new paradigm and methods that fit squarely in the historical tradition of scientific laws that open broadly generalized imaginative domains. In principle, it is best to begin measurement research from a theoretical construct model that defines the regularity sought, and informs item writing and the selection of the sample measured [125, 322, 335, 336, 338, 339]. On the other hand, empirical scaling results estimated from data sets constructed on the basis of intuitive suppositions can often teach useful lessons [129, 304].

In addition to recognizing the value of theory, Rasch also understood anomalies as indicators of new directions for qualitatively focused investigations, giving the same example of the discovery of Neptune from aberrations in the orbit of Uranus [288, p. 10] as Kuhn mentioned in his 1961 article on the function of measurement in science [184, p. 205]. Rasch's awareness of the simultaneous roles of formal ideals and concrete data in thinking was further complemented by his [291] projection that instrument equating methods will eventually lead to the deployment of abstract metrological standards and "an instrumentarium with which many kinds of problems in the social sciences can be formulated and handled."

Separating and balancing these theoretical, instrumental, and experimental purposes requires closer attention to the parts and wholes of research. Efforts in this vein, explicating the inner workings of quantitative methods, were begun in the 1920s by Thurstone [321], continued in the 1940s to 1980s by Loevinger [205, 206], Guttman [150–152], Rasch [288–290, 293], Wright [346–348, 350–352, 354–356], Luce and Tukey [208], Luce [207], and their colleagues and students [7–9, 11, 29, 90, 95, 126, 220, 246, 295, 335].

This body of work explicitly pursues the questions raised by Shiffman and Shawar [300] concerning how the global health metrics system might work more effectively to advance human wellbeing. Writing in *The Lancet*, Shiffman and Shawar echo the challenge posed by Power [282, p. 778] as to the need for social scientists to "open up the black box of performance measurement systems, to de-naturalize them and to recover the social and political work that has gone into their construction as instruments of control." Power, and Shiffman and Shawar, offer highly sophisticated perspectives on social theory, history, and philosophy in their critique of health care performance measurement systems, citing many of the same writers as those mentioned or implicated in this chapter, such as Bowker and Star, Foucault, Hacking, Heilbron, Kuhn, Miller and O'Leary, Popper, Porter, Scott, and Wise.

Instead of assuming this work of recovering the social and political construction of measurement systems has never been done, Power, and Shiffman and Shawar, might have invested more effort in seeking out and understanding efforts that have been underway for decades. One such work published in *Lancet Neurology* [162] provides an excellent explication of the technical issues, and cites a number of papers concerning the controversies involved (among others, see [11, 98, 231, 232, 351, 352, 354]. One of the papers [98] cited in Hobart et al.'s *Lancet Neurology* article explicitly concerns Power's [282] themes concerning the de-naturalization of measurement systems and the recovery of the social and political work invested in their construction. Further searches of the measurement literature would reveal other contributions developing that theme and citing many of the same sources as Shiffman and Shawar, and Power [99, 123, 124, 211–213, 215, 226, 241–244, 306].

Though there surely remain many issues in need of much more extensive elaboration and conclusive resolution, this body of work suggests that, contrary to Power's and Shiffman and Shawar's broad categorizations, performance measurement systems in health care are not merely instruments of control, but offer the possibility of extending everyday language in creative and empowering ways. Meanings are not, after all, mere ancillary implications attached to words and numbers; they are fulfilled and operationalized only in use. The question is not one of how best to adapt to inevitably reductionist homogenizations of differences but is instead a matter of learning how to navigate the complex emergent flows of meaningful and co-evolving relationships. Numbers, like all words, have a long history of being used to oppress, manipulate, control, minimize, and render invisible situations, events, or people who fall outside the boundaries of rigidly defined norms. But numbers and words also have many instances in which they serve as media connecting things and ideas in fluid and adaptable ways. Because it incorporates multilevel complexities, taking language as a model for creating new vehicles of thought offers opportunities for opening imaginations to new possibilities for negotiating local meanings and agreements on the fly in the moment of use.

Habituating end users to new technologies and complex processes need not be inherently difficult, though arriving at intuitively accessible designs will likely be very challenging. Simplifications have, after all, made routine use of extremely technical mechanical and computational devices a fact of everyday life. Motivations for producing them followed from the same economies of thought extended into larger domains as those proposed here for quality improvements in person-centered health care. Might it be possible to imagine, orchestrate, and choreograph resonant correspondences of property rights, financial standards, regulatory approvals, contract law, communications, and technical requirements to fulfill hopes and dreams for creating trustworthy, brilliant systems everyday people can use to produce exceptional results?

12.6 Conclusion

Society expresses its standards for fair dealing in the uses it makes of measurements. Measurement is at once highly technical, mathematically challenging, and a normal routine of life performed dozens of times a day. It resides in the background as something taken for granted and usually understood only to the superficial extent needed to meet basic needs in the kitchen, the workshop, the construction site, and in scheduling meetings or making travel plans.

Health care, like many other areas of contemporary life, lacks *systems* of signs and symbols for representing meaningful amounts of changes recorded categorically as present or absent, as failing or good, or as ratings on a numeric scale. Existing categorical representations allow the same signs and symbols to mean different things at different times and places when the same measured construct could plausibly be the object of interest. Not having any way of being clearly communicated, the products of the health care industry – health, functionality, abilities, and quality of care – must necessarily then lack common definitions, making it impossible to make price comparisons or to take systematic steps toward progressive improvements in outcomes. But improved approaches to thinking about and doing measurement have long been available, as also has been documentary evidence of the multiple roles measurements play in creating coherent social ecologies and economies.

The challenges are immense, as is evident in the multiple demands for operationalizing the technical issues involved in measurement modeling, instrument design and calibration, metrological traceability, the distribution of standardized forms of information throughout networks of actors, and the coordination across multidisciplinary alliances and obligatory passage points of simultaneously formal, abstract, and concrete boundary objects. But shifting the paradigm away from the unexamined and taken-for-granted modern metaphysics of centrally planned statistical solutions applied from the top down, toward emergent, distributed, and reproducible measurement solutions sets the stage for envisioning alternative outcomes for the problems encountered in the shift to quality over the last 20 years [26]. Using advanced measurement in an unmodern paradigm of actor networks might make it possible to create new approaches to sustainable change [180] that organize new visions, plans, resources, incentives, and skills capable of facilitating a new array of outcomes:

- Wholesale, systemic improvements in quality of care fulfilling their potential for being brought to scale.
- Improvements that are generalized across boundaries because theory and practice inform each other.
- Quality improvements ease financial pressures.
- Controlling costs by improving quality in “lean” processes works in health care as it does in other industries.
- Value-based pay-for-performance schemes completely replace the fee-for-service model.
- Accountability for outcomes cultivated from replicable, trusted relationships inspires positive impacts on clinician morale, and leads to clear progress in quality and safety.
- A balance between the critical need for accountability, on the one hand, and supports for a culture of trust focused on growth and learning, on the other, is brought within reach, and is grasped.
- New payment models setting more constructive priorities are political imperatives because of the way cost controls and higher profits go hand-in-hand with quality improvements.

The premise of this chapter, and this book, is that positive, constructive, and productive paths forward in health care quality improvement are indeed possible. This book develops the idea that measurement systems are not primarily mechanisms of control and domination but serve as vehicles of thought, as the media through which creative innovations find expression. Science extends everyday language such that conceptual frameworks contextualize abstract standards for representing patterns projected by individuals, which in turn contextualize observations that never conform exactly to the model.

The pragmatic idealism of integrated formalisms, abstractions, and concrete data is as old as the birth of philosophy in Plato's distinction between name and concept, figure and meaning [136, p. 100]. Plato redefined the elements of geometry to make this obvious, saying a point is an indivisible line, a line is length without breadth, etc. [46, p. 25]. This made irrational and incommensurable line segments equivalent to rational and commensurate ones; the existential threat to the Pythagorean worldview's sense of the universe as mathematical was then removed. Though Plato is often mistakenly associated with an overriding emphasis on ideal forms, his idealism is quintessentially pragmatic in accepting in his philosophy the necessary and mutually complementary roles of both abstract representations and concrete phenomena that never fit the model.

The modern metaphysics in health care quality improvement is Pythagorean in the sense of mistaking numbers and numeric relationships for existence and reality [136, p. 35]. This confusion makes it impossible to think of and act on real situations collectively, as communities of practice, because of the lack of shared languages. Mistaking numbers for quantities is akin to not seeing the forest for the trees, or confusing the map for the territory [117]. Taking an unmodern perspective advancing a metrological agenda, in contrast, operationalizes the ontological distinction between ideas and what becomes [98, 104–106, 110]. Creating useful models at abstract and formal levels of complexity distinguishes forests from the actual trees, and draws maps with features that are not in one-to-one correspondence with the concrete world. These distinctions are essential if we are to create contexts for manifesting "the world of ideas from which science is derived and which alone makes science possible" [136, p. 35].

Whitehead saw that getting from things to ideas systematically at a societal level of organization requires capacities for apprehending higher order levels of complexity and packaging them in easy-to-use technologies. As he [331, p. 61] said, "Civilization advances by extending the number of important operations which we can perform without thinking about them." The model of hierarchical complexity [60, 65, 73, 75, 76] spells out how the simplified packaging of complex operations happens; examples from the history of governance and science are instructive as to how today's transitions might be negotiated [61–63, 66].

The theory of hierarchical complexity specifies how hidden assumptions informing operations at one level of complexity become objects of operations at the next level. A child able to say, "Nice cat," for instance, may not know the alphabet or what a word is. When the child is able to say, "I know how to spell 'cat:' C-A-T," a transition from the concrete to the abstract levels of complexity has occurred. The child previously unaware language was being used now is overtly conscious of it. To Whitehead's point, however, the child had no input at all into determining the shape of the letters, their organization into a word, their pronunciation, etc. The child does not need to invest any effort in inventing communicative operations because the economy of thought facilitated by language has already done that work. That work ensures the child can inhabit a space in a community possessing shared standards, and take ownership of a trusted medium of expression with proven functionality.

As a society, we are similarly transitioning from concrete articulations focused on numeric counts to higher order quantitative abstractions, formalisms, systems, metasystems, and paradigms [74, 115, 117]. Over the last 100 years or so, dating from Thurstone's [321] work in the 1920s, measurement operations involving abstract and formal expressions of quantity could not be performed without thinking about them. Technical skill has been essential to organizing concrete observations recorded as ordinal numbers and using them as a basis for estimating abstract quantities. Those lacking access to that skill, and who could not visualize, plan, resource, or incentivize it, have been unable to perform the operations of quantification.

The emergence of metrological theory and methods marks the beginning of the consolidation and integration of a new level of hierarchical complexity in society's semiotic sociocognitive infrastructures. Metrology advances civilization by extending the number of measurement operations that can be performed by people lacking technical skill in those areas. The singularly important point elaborated in different ways by the chapters in this book concerns how numeric patterns cohere in abstract forms enabling the expression of the collectively-projected information that must be fed back via measurement to management. Quality improvement methods and lean thinking cannot fulfill their potential in health care unless the measurements used consistently stand for objectively reproducible quantities. Probabilistic models for measurement provide the analytic means by which such patterns can be identified and put to work in metrological networks of quality assured instrumentation. The chapters in this book articulate in detail the mathematics, the organization, and the interpretation of ways in which researchers and practitioners engage in that work to cultivate trust.

The consequences of making theories and data metrologically contestable must lead toward a radical departure from the current situation of incommensurable perspectives left isolated in their own domains, subject only to the competing interests of others who might command more persuasive rhetoric, greater personal charm, or larger funds. Marshaling predictive theory, explanatory models, instrument designs, consensus processes, quality assurance results, fit for purpose tolerances, traceability to standards, regulatory compliance, contractual obligations, etc. in trust markets with much higher stakes will require vastly different capacities for envisioning, planning, staffing, resourcing, and incentivizing sustainable change (taking up the frame of reference described by [180]). The needed shift in thinking does not mitigate the fact that, as was suggested some years ago [149], given the longstanding availability of the conceptual and practical resources needed for achieving fundamental measurement, continuing to accept today's dysfunctional and counterproductive ordinal quality metrics constitutes a fraudulent malpractice and major liability compromising the integrity of health care institutions.

The goal of mediating conceptual ideals and real things in thoughtful, mindful uses of words (including number words) is to create opportunities for theoretical defensibility, communications standards, and residual anomalies to complement and augment each other, with each arising in safe and productive times and places as objects of debate and actionable considerations. When these multiple purposes are accomplished via somewhat divergent and somewhat convergent ideas and methods by communities of allied participants, modern metaphysics' fixation with transcendent universals is overcome. This overcoming is not a matter of metaphysical assumptions being negated, defeated, or abandoned, since, in overcoming modern metaphysics, we must inevitably take it up and use it [80, pp. 280–281; 81, p. 62; 135, p. 240; 138, pp. 164; 184–185; 158, p. 19; 159, pp. 84–110]. This usage occurs in contexts that formally, systematically, metasystematically, and paradigmatically integrate semiotic levels of complexity. Simultaneously convergent and divergent, willed and unwilled, general and specific, global and local boundary objects already implicitly enacted across domains must become explicitly articulated in metrological research and practice. Beyond mere hope, semiotic models of language as the vehicle of thought offer actionable, pragmatic programs for sustainable change in health care's person-centered quality improvement efforts.

References

1. J.R. Ackermann, *Data, Instruments, and Theory: A Dialectical Approach to Understanding Science* (Princeton University Press, 1985)
2. Z.J. Acs, S. Estrin, T. Mickiewicz, L. Szerb, Entrepreneurship, institutional economics, and economic growth: An ecosystem perspective. Small Bus. Econ. **51**(2), 501–514 (2018)
3. A. Akmal, R. Greatbanks, J. Foote, Lean thinking in healthcare-findings from a systematic literature network and bibliometric analysis. Health Policy **124**(6), 615–627 (2020)
4. S. Alagumalai, D.D. Durtis, N. Hungi, *Applied Rasch Measurement: A Book of Exemplars* (Springer-Kluwer, 2005)
5. K. Alder, *The Measure of All Things: The Seven-Year Odyssey and Hidden Error that Transformed the World* (The Free Press, 2002)
6. R. Alsop, N. Heinsohn, *Measuring Empowerment in Practice: Structuring Analysis and Framing Indicators*, Tech. Rep. No. World Bank Policy Research Working Paper 3510 (The World Bank, 2005), p. 123
7. E.B. Andersen, Sufficient statistics and latent trait models. Psychometrika **42**(1), 69–81 (1977)
8. D. Andrich, A rating formulation for ordered response categories. Psychometrika **43**(4), 561–573 (1978)
9. D. Andrich, *Sage University Paper Series on Quantitative Applications in the Social Sciences. Vol. Series No. 07-068: Rasch Models for Measurement* (Sage, 1988)
10. D. Andrich, Distinctions between assumptions and requirements in measurement in the social sciences, in *Mathematical and Theoretical Systems: Proceedings of the 24th International Congress of Psychology of the International Union of Psychological Science*, ed. by J. A. Keats, R. Taft, R. A. Heath, S. H. Lovibond, vol. 4, (Elsevier Science Publishers, 1989), pp. 7–16
11. D. Andrich, Controversy and the Rasch model: A characteristic of incompatible paradigms? Med. Care **42**(1), I–7–I–16 (2004)
12. D. Andrich, Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. Psychometrika **75**(2), 292–308 (2010)
13. D. Andrich, I. Marais, *A Course in Rasch Measurement Theory: Measuring in the Educational, Social, and Health Sciences* (Springer, 2019)

14. W.J. Ashworth, Metrology and the state: Science, revenue, and commerce. Science **306**(5700), 1314–1317 (2004)
15. Y. Attali, Automatic item generation unleashed: An evaluation of a large-scale deployment of item models, in *International Conference on Artificial Intelligence in Education*, (Springer, 2018), pp. 17–29
16. E. Banks, The philosophical roots of Ernst Mach's economy of thought. Synthese **139**(1), 23–53 (2004)
17. S.P. Barbic, S.J. Cano, S. Mathias, The problem of patient – Centred outcome measurement in psychiatry: Why metrology hasn't mattered and why it should. J. Phys. Conf. Ser. **1044**, 012069 (2018)
18. S. Barbic, S.J. Cano, K. Tee, S. Mathias, Patient-centered outcome measurement in psychiatry: How metrology can optimize health services and outcomes, in *TMQ_Techniques, Methodologies and Quality, 10*, Special Issue on Health Metrology, (2019), pp. 10–19
19. M. Barney, W.P. Fisher Jr., Adaptive measurement and assessment. Annu. Rev. Organ. Psych. Organ. Behav. **3**, 469–490 (2016)
20. A. Barry, N. Thrift, Gabriel Tarde: Imitation, invention and economy [introduction to a special issue on G. Tarde]. Econ. Soc. **36**(4), 509–525 (2007)
21. Y. Barzel, Measurement costs and the organization of markets. J. Law Econ. **25**, 27–48 (1982)
22. P. Bech, *Rating Scales for Psychopathology, Health Status, and Quality of Life: A Compendium on Documentation in Accordance with the DSM-III-R and WHO Systems* (Springer, 1993)
23. M. Berg, S. Timmermans, Order and their others: On the constitution of universalities in medical work. Configurations **8**(1), 31–61 (2000)
24. B. Bernspång, A.G. Fisher, Differences between persons with right or left cerebral vascular accident on the assessment of motor and process skills. Arch. Phys. Med. Rehabil. **76**(12), 1144–1151 (1995)
25. W.J. Bernstein, *The Birth of Plenty: How the Prosperity of the Modern World Was Created* (McGraw-Hill, 2004)
26. D.M. Berwick, C.K. Cassel, The NAM and the quality of health care-inflecting a field. N. Engl. J. Med. **383**(6), 505–508 (2020)
27. D.M. Berwick, B. James, M.J. Coye, Connections between quality measurement and improvement. Med. Care **41**(1 (Suppl)), I30–I38 (2003)
28. W.R. Best, A Rasch model of the Crohn's disease activity index (CDAI): Equivalent levels of ranked attribute and continuous variable scales, in *Crohn's Disease: Etiology, Pathogenesis and Interventions (P. Chapter 5)*, ed. by J. N. Cadwallader, (Nova Science Publishers, 2008)
29. N. Bezruczko (ed.), *Rasch Measurement in Health Sciences* (JAM Press, 2005)
30. J.B. Bjorner, J.E. Ware, Using modern psychometric methods to measure health outcomes. Med. Outcomes Trust Monit. **3**, 2–3 (1998)
31. P. Black, M. Wilson, S. Yao, Road maps for learning: A guide to the navigation of learning progressions. Measur. Interdiscip. Res. Perspect. **9**, 1–52 (2011)
32. A. Blok, M. Nakazora, B.R. Winthereik, Infrastructuring environments. Sci. Cult. **25**(1), 1–22 (2016)
33. A. Blok, I. Farias, C. Roberts (eds.), *The Routledge Companion to Actor-Network Theory* (Routledge, 2020)
34. R.K. Bode, A.W. Heinemann, P. Semik, Measurement properties of the Galveston orientation and amnesia test (GOAT) and improvement patterns during inpatient rehabilitation. J. Head Trauma Rehabil. **15**(1), 637–655 (2000)
35. G.C. Bowker, Susan Leigh Star special issue. Mind Cult. Act. **22**(2), 89–91 (2015)
36. G.C. Bowker, How knowledge infrastructures learn, in *Infrastructures and Social Complexity: A Companion*, ed. by P. Harvey, C. B. Jensen, A. Morita, (Routledge, 2016), pp. 391–403
37. G. Bowker, S. Timmermans, A. E. Clarke, E. Balka (eds.), *Boundary Objects and Beyond: Working with Leigh Star* (MIT Press, 2015)
38. S. Brier, Cybersemiotics: A new foundation for transdisciplinary theory of information, cognition, meaningful communication and the interaction between nature and culture. Integr. Rev. **9**(2), 220–263 (2013)

39. S. Brier, Can biosemiotics be a “science” if its purpose is to be a bridge between the natural, social and human sciences? Prog. Biophys. Mol. Biol. **119**(3), 576–587 (2015)
40. T. Broer, A.P. Nieboer, R.A. Bal, Opening the black box of quality improvement collaboratives: An actor-network theory approach. BMC Health Serv. Res. **10**(1), 1–9 (2010)
41. J.C. Brown, J.E. Goldstein, T.L. Chan, R. Massof, P. Ramulu, Low Vision Research Network Study Group, Characterizing functional complaints in patients seeking outpatient low-vision services in the United States. Ophthalmology **121**(8), 1655–1662 (2014)
42. N. Bruce, D. Pope, D. Stanistreet, *Quantitative Methods for Health Research: A Practical Interactive Guide to Epidemiology and Statistics* (Wiley, 2018)
43. E.A. Burtt, *The Metaphysical Foundations of Modern Physical Science [First Edition Published in 1924]*, Rev edn. (Doubleday Anchor, 1954/1932)
44. J. Butel, K.L. Braun, The role of collective efficacy in reducing health disparities: A systematic review. Fam. Community Health **42**(1), 8–19 (2019)
45. H. Butterfield, *The Origins of Modern Science*, Rev edn. (The Free Press, 1957)
46. F. Cajori, *A History of Mathematics* (Chelsea Publishing Co, 1985)
47. M. Callon, Four models for the dynamics of science, in *Handbook of Science and Technology Studies*, ed. by S. Jasanoff, G. E. Markle, J. C. Petersen, T. Pinch, (Sage, 1995), pp. 29–63
48. M. Candea (ed.), *Routledge Advances in Sociology. Vol. 166: The Social After Gabriel Tarde: Debates and Assessments* (Routledge, 2010)
49. S. Cano, J. Hobart, R. Fitzpatrick, K. Bhatia, A. Thompson, T. Warner, Patient-based outcomes of cervical dystonia: A review of rating scales. Mov. Disord. **19**(9), 1054–1059 (2004)
50. S. Cano, L. Pendrill, S. Barbic, W.P. Fisher Jr., Patient-centred outcome metrology for healthcare decision-making. J. Phys. Conf. Ser. **1044**, 012057 (2018)
51. S. Cano, L. Pendrill, J. Melin, W.P. Fisher Jr., Towards consensus measurement standards for patient-centered outcomes. Measurement **141**, 62–69 (2019)
52. D.F. Cella, S.R. Lloyd, B.D. Wright, Cross-cultural instrument equating: Current research and future directions, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, ed. by B. Spilker, 2nd edn., (Lippincott-Raven, 1996), pp. 707–715
53. T.L. Chan, M.S. Perlmutter, M. Andrews, J.S. Sunness, J.E. Goldstein, R.W. Massof, Low Vision Research Network (LOVRNET) Study Group, Equating visual function scales to facilitate reporting of Medicare functional g-code severity/complexity modifiers for low-vision patients. Arch. Phys. Med. Rehabil. **96**(10), 1859–1865 (2015)
54. W.-C. Chang, C. Chan, Rasch analysis for outcomes measures: Some methodological considerations. Arch. Phys. Med. Rehabil. **76**(10), 934–939 (1995)
55. T.-W. Chien, W.-C. Wang, H.-Y. Wang, H.-J. Lin, Online assessment of patients' views on hospital performances using Rasch model's KIDMAP diagram. BMC Health Serv. Res. **9**, 135 (2009)
56. T.-W. Chien, J.M. Linacre, W.-C. Wang, Examining student ability using KIDMAP fit statistics of Rasch analysis in excel, in *Communications in Computer and Information Science: Vol. 201. Advances in Information Technology and Education, CSE 2011 Qingdao, China Proceedings, Part I*, ed. by H. Tan, M. Zhou, (Springer, 2011), pp. 578–585
57. T.W. Chien, Y. Chang, K.S. Wen, Y.H. Uen, Using graphical representations to enhance the quality-of-care for colorectal cancer patients. Eur. J. Cancer Care **27**(1), e12591 (2018)
58. N. Cliff, Abstract measurement theory and the revolution that never happened. Psychol. Sci. **3**, 186–190 (1992)
59. J. Cohen, The earth is round ($p < 0.05$). Am. Psychol. **49**, 997–1003 (1994)
60. M.L. Commons, Introduction to the model of hierarchical complexity and its relation to postformal structures. World Fut. J. New Paradig. Res. **64**, 305–320 (2008)
61. M.L. Commons, L.M. Bresette, Illuminating major creative scientific innovators with postformal stages, in *Handbook of Adult Development and Learning*, ed. by C. Hoare, (Oxford University Press, 2006), pp. 255–280
62. M.L. Commons, T.Q. Duong, Understanding terrorism: A behavioral developmental approach. Ethics Med. Public Health **8**, 74–96 (2019)

63. M.L. Commons, E.A. Goodheart, Consider stages of development in preventing terrorism: Does government building fail and terrorism result when developmental stages of governance are skipped? J. Adult Dev. **14**, 91–111 (2007)
64. M.L. Commons, F.A. Richards, Four postformal stages, in *Handbook of Adult Development*, ed. by J. Demick, C. Andreoletti, (Plenum Press, 2002), pp. 199–219
65. M.L. Commons, E.A. Goodheart, A. Pekker, T.L. Dawson-Tunik, K.M. Adams, Using Rasch scaled stage scores to validate orders of hierarchical complexity of balance beam task sequences. J. Appl. Meas. **9**, 182–199 (2008)
66. M.L. Commons, S.N. Ross, L.M. Bresette, The connection between postformal thought, stage transition, persistence, and ambition and major scientific innovations, in *The Oxford Handbook of Reciprocal Adult Development and Learning*, ed. by D. Artistico, J. Berry, J. Black, D. Cervone, C. Lee, H. Orom, (Oxford University Press, 2011), pp. 287–301
67. M.H. Cooper, Measure for measure? Commensuration, commodification, and metrology in emissions markets and beyond. Environ Plan A **47**(9), 1787–1804 (2015)
68. M. Cornel, R.A. Knibbe, W.M. van Zutphen, M.J. Drop, Problem drinking in a general practice population: The construction of an interval scale for severity of problem drinking. J. Stud. Alcohol **55**(4), 466–470 (1994)
69. M.J. Coye, No Toyotas in health care: Why medical care has not evolved to meet patients' needs. Health Aff. **20**(6), 44–56 (2001)
70. T.M. Cruz, The social life of biomedical data: Capturing, obscuring, and envisioning care in the digital safety-net. Soc. Sci. Med. **114670** (2021)
71. M. Danesi, Semiotics as a metalanguage for the sciences, in *Semiotics and Its Masters*, ed. by K. Bankov, P. Cobley, (DeGruyter, 2017), pp. 61–81
72. A.M. Davis, A.V. Perruccio, M. Canizares, A. Tennant, G.A. Hawker, P.G. Conaghan, E.M. Roos, J.M. Jordan, J.-F. Maillefert, M. Dougados, L.S. Lohmander, The development of a short measure of physical function for hip OA HOOS-physical function Shortform (HOOS-PS): An OARSI/OMERACT initiative. Osteoarthr. Cartil. **16**(5), 551–559 (2008)
73. T.L. Dawson, Assessing intellectual development: Three approaches, one sequence. J. Adult Dev. **11**(2), 71–85 (2004)
74. T.L. Dawson, K.W. Fischer, Z. Stein, Reconsidering qualitative and quantitative research approaches: A cognitive developmental perspective. New Ideas Psychol. **24**, 229–239 (2006)
75. T.L. Dawson, E.A. Goodheart, K. Draney, M. Wilson, M.L. Commons, Concrete, abstract, formal, and systematic operations as observed in a "Piagetian" balance-beam task, in *Advances in Rasch Measurement*, ed. by M. Garner, G. Engelhard Jr., W. P. Fisher Jr., M. Wilson, vol. 1, (JAM Press, 2010), pp. 572–590
76. T.L. Dawson-Tunik, M. Commons, M. Wilson, K. Fischer, The shape of development. Eur. J. Dev. Psychol. **2**, 163–196 (2005)
77. P. De Boeck, M. Wilson (eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. Statistics for Social and Behavioral Sciences* (Springer, 2004)
78. H. De Soto, *The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else* (Basic Books, 2000)
79. J. Derrida, *Of grammatology (G. C. Spivak, Trans.)* (Johns Hopkins University Press, 1976)
80. J. Derrida, Structure, sign and play in the discourse of the human sciences, in *Writing and Difference*, (University of Chicago Press, 1978), pp. 278–293
81. J. Derrida, Interview on writing, in *Critical Intellectuals on Writing*, ed. by G. A. Olson, L. Worsham, (State University of New York Press, 2003), pp. 61–69
82. A. Deutsch, A.W. Heinemann, K.F. Cook, L. Foster, A. Miskovic, A. Goldsmith, D. Cella, Inpatient rehabilitation quality of care from the patient's perspective: Effect of data collection timing and patient characteristics. Arch. Phys. Med. Rehabil. **100**(6), 1032–1041 (2019)
83. J. Dewey, *The Public and Its Problems* (Swallow Press, Ohio University Press, 1954)
84. J. Dewey, in *Unmodern Philosophy and Modern Philosophy*, ed. by P. Deen, (Southern Illinois University Press, 2012)

85. S. Donetto, C. Chapman, S. Brearley, A.M. Rafferty, D. Allen, G. Robert, Exploring the impact of patient experience data in acute NHS hospital trusts in England: Using actor-network theory to optimise organisational strategies and practices for improving patients' experiences of care. Health Serv. Deliv. Res. **14**(156) (2019)
86. O.D. Duncan, *Notes on Social Measurement: Historical and Critical* (Russell Sage Foundation, 1984)
87. A.K. Edgerton, Learning from standards deviations: Three dimensions for building education policies that last. Am. Educ. Res. J. **57**(4), 1525–1566 (2020)
88. S.E. Embretson, Generating items during testing: Psychometric issues and models. Psychometrika **64**(4), 407–433 (1999)
89. S.E. Embretson, *Measuring Psychological Constructs: Advances in Model-Based Approaches* (American Psychological Association, 2010)
90. G. Engelhard Jr., *Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences* (Routledge Academic, 2012)
91. E.K. Fenwick, B.S. Loe, J. Khadka, R.E. Man, G. Rees, E.L. Lamoureux, Optimizing measurement of vision-related quality of life: A computerized adaptive test for the impact of vision impairment questionnaire (IVI-CAT). Qual. Life Res. **29**(3), 765–774 (2020)
92. M.C. Ferreira, *Statistical Methods for a Comparative Study on Health Metrology. TMQ_Techniques, Methodologies and Quality, 10*, Special Issue on Health Metrology (2019), pp. 85–95
93. K. Fiedler, What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. Perspect. Psychol. Sci. **12**(1), 46–61 (2017)
94. G.H. Fischer, The linear logistic test model as an instrument in educational research. Acta Psychol. **37**, 359–374 (1973)
95. G.H. Fischer, On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. Psychometrika **46**(1), 59–77 (1981)
96. K.W. Fischer, M.J. Farrar, Generalizations about generalization: How a theory of skill development explains both generality and specificity. Int. J. Psychol. **22**(5–6), 643–677 (1987)
97. A.G. Fisher, The assessment of IADL motor skills: An application of many-faceted Rasch analysis. Am. J. Occup. Ther. **47**(4), 319–329 (1993)
98. W.P. Fisher Jr., Objectivity in measurement: A philosophical history of Rasch's separability theorem, in *Objective Measurement: Theory into Practice*, ed. by M. Wilson, vol. I, (Ablex Publishing Corporation, 1992), pp. 29–58
99. W.P. Fisher Jr., The Rasch debate: Validity and revolution in educational measurement, in *Objective Measurement: Theory into Practice*, ed. by M. Wilson, vol. II, (Ablex Publishing Corporation, 1994), pp. 36–72
100. W.P. Fisher Jr., What scale-free measurement means to health outcomes research. Phys. Med. Rehabil. State Art Rev. **11**(2), 357–373 (1997)
101. W.P. Fisher Jr., A research program for accountable and patient-centered health status measures. J. Outcome Meas. **2**(3), 222–239 (1998)
102. W.P. Fisher Jr., Foundations for health status metrology: The stability of MOS SF-36 PF-10 calibrations across samples. J. La State Med. Soc. **151**(11), 566–578 (1999)
103. W.P. Fisher Jr., Objectivity in psychosocial measurement: What, why, how. J. Outcome Meas. **4**(2), 527–563 (2000)
104. W.P. Fisher Jr., The mathematical metaphysics of measurement and metrology: Towards meaningful quantification in the human sciences, in *Renascent Pragmatism: Studies in Law and Social Science*, ed. by A. Morales, (Ashgate Publishing Co, 2003a), pp. 118–153
105. W.P. Fisher Jr., Mathematics, measurement, metaphor, metaphysics: Parts I & II. Theory Psychol. **13**(6), 753–828 (2003b)
106. W.P. Fisher Jr., Meaning and method in the social sciences. Hum. Stud. J. Philos. Soc. Sci. **27**(4), 429–454 (2004)
107. W.P. Fisher Jr., Vanishing tricks and intellectualist condescension: Measurement, metrology, and the advancement of science. Rasch measurement. Transactions **21**(3), 1118–1121 (2007) http://www.rasch.org/rmt/rmt213c.htm

108. W.P. Fisher Jr., Invariance and traceability for measures of human, social, and natural capital: Theory and application. Measurement **42**(9), 1278–1287 (2009a)
109. W.P. Fisher Jr., *NIST Critical National Need Idea White paper: Metrological Infrastructure for Human, Social, and Natural Capital* (National Institute for Standards and Technology, Washington, DC, 2009b) Tech. Rep. No. http://www.nist.gov/tip/wp/pswp/upload/202_metrological_infrastructure_for_human_social_natural.pdf. (11 p)
110. W.P. Fisher Jr., Reducible or irreducible? Mathematical reasoning and the ontological method, in *Advances in Rasch Measurement*, ed. by M. Garner, G. Engelhard Jr., W. P. Fisher Jr., M. Wilson, vol. 1, (JAM Press, 2010), pp. 12–44
111. W.P. Fisher Jr., Measure and manage: Intangible assets metric standards for sustainability, in *Business Administration Education: Changes in Management and Leadership Strategies*, ed. by J. Marques, S. Dhiman, S. Holt, (Palgrave Macmillan, 2012a), pp. 43–63
112. W.P. Fisher Jr., What the world needs now: A bold plan for new standards [third place, 2011 NIST/SES world standards day paper competition]. Stand. Eng. **64**(3), 1–3–5 (2012b) http://ssrn.com/abstract=2083975
113. W.P. Fisher Jr., Imagining education tailored to assessment as, for, and of learning: Theory, standards, and quality improvement. Assess. Learn. **2**, 6–22 (2013)
114. W.P. Fisher Jr., *A Nondualist Social Ethic: Fusing Subject and Object Horizons in Measurement. TMQ – Techniques, Methodologies, and Quality, 10*, Special Issue on Health Metrology (2019), pp. 21–40
115. W.P. Fisher Jr., Contextualizing sustainable development metric standards: Imagining new entrepreneurial possibilities. Sustainability **12**(9661), 1–22 (2020a)
116. W.P. Fisher Jr., [Entry on] Wright, Benjamin D, in *SAGE Research Methods Foundations*, ed. by P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, R. Williams, (Sage, 2020b) https://methods.sagepub.com/foundations/wright-benjamin-d
117. W.P. Fisher Jr., Bateson and Wright on number and quantity: How to not separate thinking from its relational context. Symmetry **13**(1415) (2021a)
118. W.P. Fisher Jr., Measurement as a geometry of chance and experience. Measur. Sens. **18**, 100130 (2021b)
119. W.P. Fisher Jr., Separation theorems in econometrics and psychometrics: Rasch, Frisch, two fishers, and implications for measurement. J. Interdiscip. Econ. **OnlineFirst**, 1–32 (2021c)
120. W.P. Fisher Jr., A.J. Stenner, Integrating qualitative and quantitative research approaches via the phenomenological method. Int. J. Multi. Res. Approac. **5**(1), 89–103 (2011a)
121. W.P. Fisher Jr., A.J. Stenner, A technology roadmap for intangible assets metrology, in *Fundamentals of Measurement Science. International Measurement Confederation (IMEKO) TC1-TC7-TC13 Joint Symposium, Jena, Germany, September 2*, (2011b) http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24493/ilm1-2011imeko-018.pdf
122. W.P. Fisher Jr., A.J. Stenner, On the potential for improved measurement in the human and social sciences, in *Pacific Rim Objective Measurement Symposium 2012 Conference Proceedings*, ed. by Q. Zhang, H. Yang, (Springer, 2013), pp. 1–11
123. W.P. Fisher Jr., A.J. Stenner, Theory-based metrological traceability in education: A reading measurement network. Measurement **92**, 489–496 (2016)
124. W.P. Fisher Jr., M. Wilson, Building a productive trading zone in educational assessment research and practice. Pensamiento Educativo: Revista de Investigacion Educacional Latinoamericana **52**(2), 55–78 (2015)
125. W.P. Fisher Jr., M. Wilson, An online platform for sociocognitive metrology: The BEAR assessment system software. Meas. Sci. Technol. **31**(034006) (2020)
126. W. P. Fisher Jr., B. D. Wright (eds.), Applications of probabilistic conjoint measurement. Int. J. Educ. Res. **21**(6), 557–664 (1994)
127. W.P. Fisher Jr., R.F. Harvey, K.M. Kilgore, New developments in functional assessment: Probabilistic models for gold standards. NeuroRehabilitation **5**(1), 3–25 (1995)
128. W.P. Fisher Jr., R.F. Harvey, P. Taylor, K.M. Kilgore, C.K. Kelly, Rehabits: A common language of functional assessment. Arch. Phys. Med. Rehabil. **76**(2), 113–122 (1995)

129. W.P. Fisher Jr., J. Melin, C. Möller, *Metrology for Climate-Neutral Cities (RISE Research Institutes of Sweden AB No. RISE report 2021:84)* (RISE, Gothenburg, 2021) http://ri.diva-portal.org/smash/record.jsf?pid=diva2%3A1616048&dswid=-7140 (79 p)
130. W.P. Fisher Jr., E.P.-T. Oon, S. Benson, Rethinking the role of educational assessment in classroom communities: How can design thinking address the problems of coherence and complexity? Educ. Des. Res. **5**(1), 1–33 (2021)
131. M.J.C. Forgeard, E. Jayawickreme, M. Kern, M.E.P. Seligman, Doing the right thing: Measuring wellbeing for public policy. Int. J. Wellbeing **1**(1), 79–106 (2011)
132. P.H. Fowler, J. Craig, L.D. Fredendall, U. Damali, Perioperative workflow: Barriers to efficiency, risks, and satisfaction. AORN J. **87**(1), 187–208 (2008)
133. G. Franck, The scientific economy of attention: A novel approach to the collective rationality of science. Scientometrics **55**(1), 3–26 (2002)
134. G. Franck, The economy of attention. J. Sociol. **55**(1), 8–19 (2019)
135. H.-G. Gadamer, *Philosophical Hermeneutics (D. E. Linge, Trans.)* (University of California Press, 1976)
136. H.-G. Gadamer, *Dialogue and Dialectic: Eight Hermeneutical Studies on Plato (P. C. Smith, Trans.)* (Yale University Press, 1980)
137. H.-G. Gadamer, *Truth and Method (J. Weinsheimer & D. G. Marshall, Trans.)*, Rev edn. (Crossroad, 1989)
138. H.-G. Gadamer, *Heidegger's Ways (D. J. Schmidt, Ed.) (J. W. Stanley, Trans.). SUNY Series in Contemporary Continental Philosophy* (SUNY Press, 1994)
139. P. Galison, *Image and Logic: A Material Culture of Microphysics* (University of Chicago Press, 1997)
140. P. Galison, Trading zone: Coordinating action and belief, in *The Science Studies Reader*, ed. by M. Biagioli, (Routledge, 1999), pp. 137–160
141. P. Galison, D.J. Stump, *The Disunity of Science: Boundaries, Contexts, and Power* (Stanford University Press, 1996)
142. M.P. Gallaher, B.R. Rowe, A.V. Rogozhin, S.A. Houghton, J.L. Davis, M.K. Lamvik, J.S. Geikler, *Economic Impact of Measurement in the Semiconductor Industry (Tech. Rep. No. 07-2)* (National Institute for Standards and Technology, Gaithersburg, 2007) 191 p
143. R. Gasché, "A certain walk to follow": Derrida and the question of method. Epoché J. Hist. Philos. **18**(2), 525–550 (2014)
144. M.J. Gierl, T.M. Haladyna, *Automatic Item Generation: Theory and Practice* (Routledge, 2012)
145. J.E. Goldstein, E. Fenwick, R.P. Finger, V. Gothwal, M.L. Jackson, E. Lamoureux, G. Rhees, R. Massof, Calibrating the impact of vision impairment (IVI): Creation of a sample-independent visual function measure for patient-centered outcomes research. Transl. Vis. Sci. Technol. **7**(6), 38 (2018)
146. P. Goodman, Emphasis on growth is called misguided. New York Times (2009). https://www.nytimes.com/2009/09/23/business/economy/23gdp.html
147. D. Greaves, Meeting the educational needs of students with learning difficulties: A sociological study of three schools in Victoria. Aust. J. Learn. Disabil. **4**(3), 12–20 (1999)
148. G. Grimby, E. Andrén, E. Holmgren, B. Wright, J.M. Linacre, V. Sundh, Structure of a combination of functional Independence measure and instrumental activity measure items in community-living persons: A study of individuals with spina bifida. Arch. Phys. Med. Rehabil. **77**(11), 1109–1114 (1996)
149. G. Grimby, A. Tennant, L. Tesio, The use of raw scores from ordinal scales: Time to end malpractice? J. Rehabil. Med. **44**, 97–98 (2012)
150. L. Guttman, A basis for scaling qualitative data. Am. Sociol. Rev. **9**, 139–150 (1944)
151. L. Guttman, What is not what in statistics. The Statistician **26**, 81–107 (1977)
152. L. Guttman, The illogic of statistical inference for cumulative science. Appl. Stoch. Models Data Anal. **1**, 3–10 (1985)
153. P. Hagell, Measuring activities of daily living in Parkinson's disease: On a road to nowhere and back again? Measurement **132**, 109–124 (2019)

154. S.M. Haley, L.H. Ludlow, Applicability of the hierarchical scales of the tufts assessment of motor performance for school-aged children and adults with disabilities. Phys. Ther. **72**(3), 191–202 (1992)
155. H. Harbers, *Inside the Politics of Technology: Agency and Normativity in the Co-Production of Technology and Society* (Amsterdam University Press, 2005)
156. T.E. Hardwicke, R.T. Thibault, J.E. Kosie, J.D. Wallach, M.C. Kidwell, J.P. Ioannidis, Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). Perspect. Psychol. Sci., 1745691620979806 (2021)
157. P.A. Heelan, Natural science as a hermeneutic of instrumentation. Philos. Sci. **50**, 181–204 (1983)
158. M. Heidegger, A dialogue on language between a Japanese and an inquirer, in *Heidegger, on the Way to Language*, ed. by M. In, (Harper & Row, 1971), pp. 1–54
159. M. Heidegger, *The End of Philosophy* (Harper & Row, 1973)
160. M. Heidegger, *The Principle of Reason (R. Lilly, Trans.)* (Indiana University Press, 1991)
161. A.W. Heinemann, W.P. Fisher Jr., R. Gershon, Improving health care quality with outcomes management. J. Prosthet. Orthot. **18**(1), 46–50 (2006)
162. J.C. Hobart, S.J. Cano, J.P. Zajicek, A.J. Thompson, Rating scales as outcome measures for clinical trials in neurology: Problems, solutions, and recommendations. Lancet Neurol. **6**, 1094–1105 (2007)
163. K. Holm, J. Kavanagh, An approach to modifying self-report instruments. Res. Nurs. Health **8**, 13–18 (1985)
164. T.A. Holster, J.W. Lake, From raw scores to Rasch in the classroom. Shiken **19**(1), 32–41 (2015)
165. I. Hong, Y. Lim, H. Han, C.C. Hay, H.S. Woo, Application of the Korean version of the modified Barthel index: Development of a keyform for use in clinical practice. Hong Kong J. Occup. Ther. **29**(1), 39–46 (2017)
166. L.F. Hornke, M.W. Habon, Rule-based item bank construction and evaluation within the linear logistic framework. Appl. Psychol. Meas. **10**(4), 369–380 (1986)
167. B.J. Hunt, The ohm is where the art is: British telegraph engineers and the development of electrical standards. Osiris Res. J. Devot. Hist. Sci. Its Cult. Influ. **9**, 48–63 (1994)
168. E. Hutchins, *Cognition in the Wild* (MIT Press, 1995)
169. E. Hutchins, Concepts in practice as sources of order. Mind Cult. Act. **19**, 314–323 (2012)
170. D. Ihde, *Instrumental Realism: The Interface between Philosophy of Science and Philosophy of Technology*, The Indiana Series in the Philosophy of Technology (Indiana University Press, 1991)
171. Institute of Medicine, Committee on quality of health Care in America, in *To Err Is Human: Building a Safer Health System*, ed. by L. Kohn, J. Corrigan, M. Donaldson, (National Academy Press, 1999)
172. Institute of Medicine, *Crossing the Quality Chasm: A New Health System for the 21st Century* (National Academy Press, 2001)
173. J.P. Ioannidis, The reproducibility wars: Successful, unsuccessful, uninterpretable, exact, conceptual, triangulated, contested replication. Clin. Chem. **63**(5), 943–945 (2017)
174. S. Jasanoff, *States of Knowledge: The Co-Production of Science and Social Order. International Library of Sociology* (Routledge, 2004)
175. T.H. Jen, T.W. Chien, W. Chou, A novel approach to the classification of performance on inpatient perception of hospitalization experience across the US 52 states. (2019). https://doi.org/10.21203/rs.2.14225/v1
176. P. Joyce (ed.), *The Social in Question: New Bearings* (Routledge, 2002)
177. H. Karasti, F. Millerand, C.M. Hine, G.C. Bowker, Knowledge infrastructures: Intro to Part I. Sci. Technol. Stud. **29**(1), 2–12 (2016)
178. P.R. Kelley, C.F. Schumacher, The Rasch model: Its use by the National Board of medical examiners. Eval. Health Prof. **7**(4), 443–454 (1984)

179. G. Kielhofner, L. Dobria, K. Forsyth, S. Basu, The construction of keyforms for obtaining instantaneous measures from the occupational performance history interview ratings scales. OTJR: Occup. Particip. Health **25**(1), 23–32 (2005)
180. T.P. Knoster, R.A. Villa, J.S. Thousand, A framework for thinking about systems change, in *Restructuring for Caring and Effective Education: Piecing the Puzzle Together*, ed. by R. A. Villa, J. S. Thousand, 2nd edn., (Paul H. Brookes, 2000), pp. 93–128
181. A. Kosh, M.A. Simpson, L. Bickel, M. Kellog, E. Sanford-Moore, A cost-benefit analysis of automatic item generation. Educ. Meas. Issues Pract. **38**(1), 48–53 (2019)
182. T.S. Kuhn, The function of measurement in modern physical science. Isis **52**(168), 161–193 (1961) (Rpt. in T. S. Kuhn, (Ed.). (1977). The essential tension (pp. 178–224). University of Chicago Press)
183. T.S. Kuhn, *The Structure of Scientific Revolutions* (University of Chicago Press, 1970)
184. T.S. Kuhn, *The Essential Tension: Selected Studies in Scientific Tradition and Change* (University of Chicago Press, 1977)
185. L. Kurunmäki, Professional vs financial capital in the field of health care-struggles for the redistribution of power and control. Acc. Organ. Soc. **24**(2), 95–124 (1999)
186. S. Labovitz, The nonutility of significance tests: The significance of tests of significance reconsidered. Pac. Sociol. Rev. **13**(3), 141–148 (1970)
187. M.A. Lapré, L.N. Van Wassenhove, Learning across lines: The secret to more efficient factories. Harv. Bus. Rev. **80**(10), 107–111 (2002)
188. B. Latour, *Science in Action: How to Follow Scientists and Engineers through Society* (Harvard University Press, 1987)
189. B. Latour, Postmodern? No, simply amodern: Steps towards an anthropology of science. Stud. Hist. Phil. Sci. **21**(1), 145–171 (1990a)
190. B. Latour, Technology is society made durable. Sociol. Rev. **38**(1-S), 103–131 (1990b)
191. B. Latour, The impact of science studies on political philosophy. Sci. Technol. Hum. Values **16**(1), 3–19 (1991)
192. B. Latour, *We Have Never Been Modern* (Harvard University Press, 1993)
193. B. Latour, To modernise or ecologise? That is the question, in *Remaking Reality: Nature at the Millennium*, ed. by B. Braun, N. Castree, (Routledge, 1998), pp. 221–242
194. B. Latour, *Politics of Nature: How to Bring the Sciences into Democracy* (Harvard University Press, 2004)
195. B. Latour, *Reassembling the Social: An Introduction to Actor-Network-Theory. Clarendon Lectures in Management Studies* (Oxford University Press, 2005)
196. B. Latour, Tarde's idea of quantification, in *The Social after Gabriel Tarde: Debates and Assessments*, ed. by M. Candea, (Routledge, 2010), pp. 145–162
197. Y.-J. Lee, O.-C. Yeoh, Kidmap construction by hand. Rasch measurement. Transactions **12**(2), 640 (1998) http://www.rasch.org/rmt/rmt122t.htm
198. T. Lenoir, Instituting science: The cultural production of scientific disciplines, in *Writing Science*, ed. by T. Lenoir, H. U. Gumbrecht, (Stanford University Press, 1997)
199. T. Lenoir, *Inscribing Science: Scientific Texts and the Materiality of Communication* (Stanford University Press, 1998)
200. J.M. Linacre, *Many-facet Rasch Measurement* (MESA Press, 1989) http://www.winsteps.com/a/facets-manual.pdf
201. J.M. Linacre, Instantaneous measurement and diagnosis. Phys. Med. Rehabil. State Art Rev. **11**(2), 315–324 (1997) http://www.rasch.org/memo60.htm
202. J.M. Linacre, Spanish-language KIDMAP. Rasch Measur. Trans. **12**(4) (1998) http://www.rasch.org/rmt/rmt124k.htm
203. J.M. Linacre, *A User's Guide to WINSTEPS Rasch-Model Computer Program, v. 5.1.1* (Winsteps.com, 2021) https://www.winsteps.com/manuals.htm
204. J. Liu, Development and translation of measurement findings for the motivation assessment for team readiness, integration, and Collaboration Self-Scoring Form. Am. J. Occup. Ther. **72**-(4_Supplement_1), 7211500015p1-7211500015p1 (2018)

205. J. Loevinger, A systematic approach to the construction and evaluation of tests of ability. Psychol. Monogr. **61**(285), 1–49 (1947)
206. J. Loevinger, Objective tests as instruments of psychological theory. Psychol. Rep. **3**, 635–694 (1957)
207. R.D. Luce, Dimensionally invariant numerical laws correspond to meaningful qualitative relations. Philos. Sci. **45**, 1–16 (1978)
208. R.D. Luce, J.W. Tukey, Simultaneous conjoint measurement: A new kind of fundamental measurement. J. Math. Psychol. **1**(1), 1–27 (1964)
209. L.H. Ludlow, S.M. Haley, B.M. Gans, A hierarchical model of functional performance in rehabilitation medicine: The tufts assessment of motor performance. Eval. Health Prof. **15**, 59–74 (1992)
210. M.E. Lunz, B.D. Wright, J.M. Linacre, Measuring the impact of judge severity on examination scores. Appl. Meas. Educ. **3**(4), 331–345 (1990)
211. M.D. Maraun, Meaning and mythology in the factor analysis model. Multivar. Behav. Res. **31**(4), 603–616 (1996a)
212. M.D. Maraun, Metaphor taken as math: Indeterminancy in the factor analysis model. Multivar. Behav. Res. **31**(4), 517–538 (1996b)
213. M. Maraun, S.M. Gabriel, Illegitimate concept equating in the partial fusion of construct validation theory and latent variable modeling. New Ideas Psychol. **31**(1), 32–42 (2013)
214. L. Mari, M. Wilson, An introduction to the Rasch measurement approach for metrologists. Measurement **51**, 315–327 (2014)
215. L. Mari, A. Maul, D.T. Irribara, M. Wilson, Quantities, quantification, and the necessary and sufficient conditions for measurement. Measurement **100**, 115–121 (2016)
216. L. Mari, M. Wilson, A. Maul, *Measurement Across the Sciences*, Springer Series in Measurement Science and Technology (Springer, 2021)
217. R.W. Massof, Evaluation of the NEI visual functioning questionnaire as an interval measure of visual ability in low vision. Vis. Res. **41**(3), 397–413 (2001)
218. R.W. Massof, The measurement of vision disability. Optom. Vis. Sci. **79**(8), 516–552 (2002)
219. R.W. Massof, Application of stochastic measurement models to visual function rating scale questionnaires. Ophthalmic Epidemiol. **12**(2), 103–124 (2005)
220. R.W. Massof, Editorial: Moving toward scientific measurements of quality of life. Ophthalmic Epidemiol. **15**, 209–211 (2008)
221. R.W. Massof, A general theoretical framework for interpreting patient-reported outcomes estimated from ordinally scaled item responses. Stat. Methods Med. Res. **23**(5), 409–429 (2014)
222. R.W. Massof, L. Ahmadian, What do different visual function questionnaires measure? Ophthalmic Epidemiol. **14**(4), 198–204 (2007)
223. R.W. Massof, C. Bradley, A strategy for measuring patient preferences to incorporate in benefit-risk assessment of new ophthalmic devices and procedures. J. Phys. Conf. Ser. **772**, 012047 (2016)
224. R.W. Massof, L. Ahmadian, L.L. Grover, J.T. Deremeik, J.E. Goldstein, C. Rainey, C. Epstein, G.D. Barnett, The activity inventory: An adaptive visual function questionnaire. Optom. Vis. Sci. **84**, 763–774 (2007)
225. G. N. Masters, J. P. Keeves (eds.), *Advances in Measurement in Educational Research and Assessment* (Pergamon, 1999)
226. J.A. McGrane, A. Maul, The human sciences, models and metrological mythology. Measurement **152**(107346) (2020)
227. R.J. Mead, The ISR: Intelligent student reports. J. Appl. Meas. **10**(2), 208–224 (2009)
228. P.E. Meehl, Theory-testing in psychology and physics: A methodological paradox. Philos. Sci. **34**(2), 103–115 (1967)
229. J. Melin, L. Pendrill, S. Cano, EMPIR NeuroMet 15HLT04 Consortium, Towards patient-centred cognition metrics. J. Phys. Conf. Ser. **1379**(012029) (2019)
230. J. Melin, S. Cano, L. Pendrill, The role of entropy in construct specification equations (CSE) to improve the validity of memory tests. Entropy **23**(2), 212 (2021)

231. C.T. Merbitz, J. Morris, J.C. Grip, Ordinal scales and the foundations of misinference. Arch. Phys. Med. Rehabil. **70**, 308–312 (1989)
232. J. Michell, Measurement scales and statistics: A clash of paradigms. Psychol. Bull. **100**, 398–407 (1986)
233. J. Michell, *Measurement in Psychology: A Critical History of a Methodological Concept* (Cambridge University Press, 1999)
234. P. Miller, T. O'Leary, Mediating instruments and making markets: Capital budgeting, science and the economy. Acc. Organ. Soc. **32**(7–8), 701–734 (2007)
235. T. Mitchell, Rethinking economy. Geoforum **39**, 1116–1121 (2008)
236. M. Molz, M.G. Edwards, Research across boundaries: Introduction to the first part of the special issue on the international symposium: Research across boundaries. Integral Rev. **9**(2), 1–11 (2013)
237. J. Morrison, W.P. Fisher Jr., Connecting learning opportunities in STEM education: Ecosystem collaborations across schools, museums, libraries, employers, and communities. J. Phys. Conf. Ser. **1065**(022009) (2018)
238. J. Morrison, W.P. Fisher Jr., Measuring for management in science, technology, engineering, and mathematics learning ecosystems. J. Phys. Conf. Ser. **1379**(012042) (2019)
239. J. Morrison, W.P. Fisher Jr., *The Measure STEM Caliper Development Initiative* [Online]. http://bearcenter.berkeley.edu/seminar/measure-stem-caliper-development-initiative-online. BEAR Seminar Series. (University of California, Berkeley, 2020)
240. J. Morrison, W.P. Fisher Jr., Caliper: Measuring success in STEM learning ecosystems. Measur. Sens. **18**, 100327 (2021)
241. S.A. Mulaik, The critique of pure statistics: Artifact and objectivity in multivariate statistics, in *Advances in Social Science Methodology*, ed. by B. Thompson, vol. 3, (JAI Press, 1993)
242. S.A. Mulaik, Kant, Wittgenstein, objectivity, and structural equation models, in *Cognitive Assessment: A Multidisciplinary Perspective*, ed. by C. R. Reynolds, (Plenum, 1994)
243. S.A. Mulaik, Factor analysis in not just a model in pure mathematics. Multivar. Behav. Res. **31**(4), 655–661 (1996a)
244. S.A. Mulaik, On Maraun's deconstructing of factor indeterminacy with constructed factors. Multivar. Behav. Res. **31**(4), 579–592 (1996b)
245. D. Narayan, *Empowerment and Poverty Reduction: A Sourcebook* (The World Bank, Washington, DC, 2002)
246. L. Narens, R.D. Luce, Measurement: The theory of numerical assignments. Psychol. Bull. **99**(2), 166–180 (1986)
247. National Institute for Standards and Technology, Appendix C: Assessment examples. Economic impacts of research in metrology, in *Assessing fundamental science: A report from the Subcommittee on Research, Committee on Fundamental Science*, ed. by C. o. F. S. Subcommittee on Research, (National Standards and Technology Council, 1996) https://wayback.archive-it.org/5902/20150628164643/http://www.nsf.gov/statistics/ostp/assess/nstcafsk.htm#Topic%207
248. National Institute for Standards and Technology. Outputs and Outcomes of NIST Laboratory Research. (2009). Retrieved 18 April 2020, from NIST: https://www.nist.gov/director/outputs-and-outcomes-nist-laboratory-research
249. N.J. Nersessian, Maxwell and "the method of physical analogy": Model-based reasoning, generic abstraction, and conceptual change, in *Reading Natural Philosophy: Essays in the History and Philosophy of Science and Mathematics*, ed. by D. Malament, (Open Court, 2002), pp. 129–166
250. N.J. Nersessian, Model-based reasoning in distributed cognitive systems. Philos. Sci. **73**, 699–709 (2006)
251. N.J. Nersessian, *Creating Scientific Concepts* (MIT Press, 2008)
252. D. Neyland, V. Ehrenstein, S. Milyaeva, Mundane market matters: On sensitive metrology and the governance of market-based interventions for global health. Rev. Fr. Sociol. **58**(3), 425–449 (2017)

253. T. Nolan, D.M. Berwick, All-or-none measurement raises the bar on performance. JAMA **295**(10), 1168–1170 (2006)
254. D.C. North, *Institutions, Institutional Change, and Economic Performance* (Cambridge University Press, 1990)
255. W. Nöth, The semiotics of models. Sign Syst. Stud. **46**(1), 7-43 (2018)
256. A. Nuzzo, The idea of 'method' in Hegel's science of logic – A method for finite thinking and absolute reason. Hegel Bull. **20**(1–2), 1–17 (1999)
257. A. Nuzzo, Thinking being: Method in Hegel's logic of being, in *A Companion to Hegel*, ed. by S. Houlgate, M. Bauer, (Blackwell, 2011), pp. 109–138
258. A. Nuzzo, *Approaching Hegel's Logic, Obliquely: Melville, Moliere, Beckett* (SUNY Press, 2018)
259. J. O'Connell, Metrology: The creation of universality by the circulation of particulars. Soc. Stud. Sci. **23**, 129–173 (1993)
260. A. Olteanu, Multimodal modeling: Bridging biosemiotics and social semiotics. Biosemiotics, 1–23 (2021)
261. R. Othman, N.A. Hashim, Typologizing organizational amnesia. Learn. Organ. **11**(3), 273–284 (2004)
262. J. Overwijk, Paradoxes of rationalisation: Openness and control in critical theory and Luhmann's systems theory. Theory Cult. Soc. **38**(1), 127–148 (2021)
263. Y.A. Ozcan, *Quantitative Methods in Health Care Management: Techniques and Applications*, vol 4 (Wiley, 2005)
264. T. Papadopoulos, Y. Merali, Stakeholder dynamics and the implementation of process innovations: The case of lean thinking in a UK NHS hospital trust. Int. J. Healthc. Technol. Manag. **10**(4–5), 303–324 (2009)
265. H.H. Pattee, Dynamic and linguistic modes of complex systems. Int. J. Gen. Syst. **3**(4), 259–266 (1977)
266. H.H. Pattee, Universal principles of measurement and language functions in evolving systems, in *Complexity, Language, and Life: Mathematical Approaches*, ed. by J. L. Casti, A. Karlqvist, (Springer Verlag, 1985), pp. 268–281
267. J.K. Peat, C. Mellis, K. Williams, W. Xuan, *Health Science Research: A Handbook of Quantitative Methods* (Routledge, 2020)
268. C.S. Peirce, *Philosophical writings of Peirce (J. Buchler, Ed.)* (Dover, 1955)
269. C.S. Peirce, *The Essential Peirce: Selected Philosophical Writings, Volume I (1867–1893) (N. Houser & C. Kloesel, Eds.)* (Indiana University Press, 1992)
270. L.R. Pendrill, Assuring measurement quality in person-centred healthcare. Measur. Sci. Technol. **29**(3), 034003 (2018)
271. L.R. Pendrill, *Quality Assured Measurement: Unification Across Social and Physical Sciences*, Springer Series in Measurement Science and Technology (Springer, 2019)
272. L. Pendrill, W.P. Fisher Jr., Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. Measurement **71**, 46–55 (2015)
273. L.R. Pendrill, J. Melin, *Measuring Counted Fractions in Healthcare. TMQ_Techniques, Methodologies and Quality, 10*, Special Issue on Health Metrology (2019), pp. 61–69
274. A.V. Perruccio, L.S. Lohmander, M. Canizares, A. Tennant, G.A. Hawker, P.G. Conaghan, E.M. Roos, J.M. Jordan, J.-F. Maillefert, M. Dougados, A.M. Davis, The development of a short measure of physical function for knee OA KOOS-physical function Shortform (KOOS-PS): An OARSI/OMERACT initiative. Osteoarthr. Cartil. **16**(5), 542–550 (2008)
275. M. Pettinari, P. Sergeant, B. Meuris, Quantification of operational learning in off-pump coronary bypass. Eur. J. Cardiothorac. Surg. **43**(4), 709–714 (2013)
276. D. Pflueger, Knowing patients: The customer survey and the changing margins of accounting in healthcare. Acc. Organ. Soc. **53**, 17–33 (2016)
277. H. Poinstingl, The linear logistic test model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. Psychol. Sci. Q. **51**, 123–134 (2009)

278. N. Poposki, N. Majcen, P. Taylor, Assessing publically financed metrology expenditure against economic parameters. Accredit. Q. Assur. J. Q. Comp. Reliab. Chem. Measur. **14**(7), 359–368 (2009)

279. M. Poppendieck, Principles of lean thinking. IT Manag. Sel. **18**, 1–7 (2011)

280. T.M. Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton University Press, 1995)

281. N. Postman, *Technopoly: The Surrender of Culture to Technology* (Vintage Books, 1992)

282. M. Power, Counting, control and calculation: Reflections on measuring and management. Hum. Relat. **57**, 765–783 (2004)

283. P.E. Prasetyo, N.R. Kistanti, Human capital, institutional economics and entrepreneurship as a driver for quality & sustainable economic growth. Entrep. Sustain. Issues **7**(4), 2575–2589 (2020)

284. B. Prien, How to predetermine the difficulty of items of examinations and standardized tests. Stud. Educ. Eval. **15**, 309–317 (1989)

285. B. Prodinger, A.A. Küçükdeveci, S. Kutlay, A.H. Elhan, S. Kreiner, A. Tennant, Cross-diagnostic scale-banking using Rasch analysis: Developing a common reference metric for generic and health condition-specific scales in people with rheumatoid arthritis and stroke. J. Rehabil. Med. **52**(10), 1–10 (2020)

286. M. Quaglia, L. Pendrill, J. Melin, S. Cano, 15HLT04 NeuroMet Consortium, *Innovative Measurements for Improved Diagnosis and Management of Neurodegenerative Diseases (EMPIR NeuroMet)* (EURAMET, Teddington, 2016–2019). https://www.lgcgroup.com/our-programmes/empir-neuromet/neuromet-landing-page/. (36 p)

287. M. Quaglia, L. Pendrill, J. Melin, S. Cano, 18HLT09 NeuroMet2 Consortium, *Publishable Summary for 18HLT09 NeuroMet2: Metrology and Innovation for Early Diagnosis and Accurate Stratification of Patients with Neurodegenerative Diseases (EMPIR NeuroMet).* (EURAMET, Teddington, 2019–2022). https://www.lgcgroup.com/our-programmes/empir-neuromet/neuromet-landing-page/. (5 p)

288. G. Rasch, *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, 1960)

289. G. Rasch, On general laws and the meaning of measurement in psychology, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Volume IV: Contributions to Biology and Problems of Medicine*, ed. by J. Neyman, (University of California Press, 1961), pp. 321–333. http://www.rasch.org/memo1960.pdf

290. G. Rasch, An individualistic approach to item analysis, in *Readings in Mathematical Social Science*, ed. by P. F. Lazarsfeld, N. W. Henry, (Science Research Associates, 1966), pp. 89–108

291. G. Rasch, Retirement Lecture of 9 March 1972: Objectivity in Social Sciences: A Method Problem (Cecilie Kreiner, Trans.). Rasch Measur. Trans. **24**(1), 1252–1272 (1972/2010). http://www.rasch.org/rmt/rmt241.pdf

292. Rasch, G., 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 Measur. Trans. **24**(4), 1309 (1973/2011). http://www.rasch.org/rmt/rmt244.pdf

293. G. Rasch, On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. Dan. Yearb. Philos. **14**, 58–94 (1977) https://www.rasch.org/memo18.htm

294. P. Ricoeur, The model of the text: Meaningful action considered as a text, in *Hermeneutics and the Human Sciences: Essays on Language, Action and Interpretation*, ed. by J. B. Thompson, (Cambridge University Press, 1981), pp. 197–221

295. T. Salzberger, *Measurement in Marketing Research: An Alternative Framework* (Edward Elgar, 2009)

296. T. Salzberger, S. Cano, L. Abetz-Webb, E. Afolalu, C. Chrea, R. Weitkunat, K. Fagerström, J. Rose, Addressing traceability in social measurement: Establishing a common metric for dependence. J. Phys. Conf. Ser. **1379**(1), 012024 (2019)
297. J.C. Scott, *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed* (Yale University Press, 1998)
298. T.A. Sebeok, *Signs: An Introduction to Semiotics* (University of Toronto Press, 2001)
299. S. Shapin, The invisible technician. Am. Sci. **77**(6), 554–563 (1989)
300. J. Shiffman, Y.R. Shawar, Strengthening accountability of the global health metrics enterprise. Lancet **395**, 1452–1456 (2020). https://doi.org/10.1016/S0140-6736(20)30416-5
301. D.A. Shore, *Launching and Leading Change Initiatives in Health Care Organizations: Managing Successful Projects*, vol 213 (Jossey-Bass, 2014)
302. K. Sijtsma, Correcting fallacies in validity, reliability, and classification. Int. J. Test. **8**(3), 167–194 (2009)
303. K. Sijtsma, Playing with data – Or how to discourage questionable research practices and stimulate researchers to do things right. Psychometrika **81**(1), 1–15 (2016)
304. B.J. Silverstein, W.P. Fisher Jr., K.M. Kilgore, R.F. Harvey, J.P. Harley, Applying psychometric criteria to functional assessment in medical rehabilitation: II. Defining interval measures. Arch. Phys. Med. Rehabil. **73**(6), 507–518 (1992)
305. R.M. Smith, P. Taylor, Equating rehabilitation outcome scales: Developing common metrics. J. Appl. Meas. **5**(3), 229–242 (2004)
306. S. Solloway, W.P. Fisher Jr., Mindfulness in measurement: Reconsidering the measurable in mindfulness. Int. J. Transpers. Stud. **26**, 58–81 (2007) http://digitalcommons.ciis.edu/ijts-transpersonalstudies/vol26/iss1/8
307. P. Sonnleitner, Using the LLTM to evaluate an item-generating system for reading comprehension. Psychol. Sci. Q. **50**(3), 345–362 (2008)
308. S.L. Star, This is not a boundary object: Reflections on the origin of a concept. Sci. Technol. Hum. Values **35**(5), 601–617 (2010)
309. S.L. Star, J.R. Griesemer, Institutional ecology, ‘translations,’ and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. Soc. Stud. Sci. **19**(3), 387–420 (1989)
310. S.L. Star, K. Ruhleder, Steps toward an ecology of infrastructure: Design and access for large information spaces. Inf. Syst. Res. **7**(1), 111–134 (1996)
311. A.J. Stenner, M. Smith III, Testing construct theories. Percept. Mot. Skills **55**, 415–426 (1982)
312. A.J. Stenner, M. Smith III, D.S. Burdick, Toward a theory of construct definition. J. Educ. Meas. **20**(4), 305–316 (1983)
313. A.J. Stenner, C. Swartz, S. Hanlon, C. Emerson, *Personalized Learning Platforms* (Presented at the Pearson Global Research Conference, Fremantle, 2012)
314. A.J. Stenner, W.P. Fisher Jr., M.H. Stone, D.S. Burdick, Causal Rasch models. Front. Psychol. Quant. Psychol. Measur. **4**(536), 1–14 (2013)
315. A.J. Stenner, M.H. Stone, W.P. Fisher Jr., The unreasonable effectiveness of theory based instrument calibration in the natural sciences: What can the social sciences learn? J. Phys. Conf. Ser. **1044**(012070) (2018)
316. D. Stokols, S. Misra, M.G. Runnerstrom, J.A. Hipp, Psychology in an age of ecological crisis: From personal angst to collective action. Am. Psychol. **64**(3), 181–193 (2009)
317. M.H. Stone, *Knox's cube test – Revised* (Stoelting, 2002)
318. J. Sutton, C.B. Harris, P.G. Keil, A.J. Barnier, The psychology of memory, extended cognition, and socially distributed remembering. Phenomenol. Cogn. Sci. **9**(4), 521–560 (2010)
319. A. Tennant, G. Grimby, C. Marincek, H. Phillips, H. Ring, F. Biering-Sorensen, L. Tesio, J.-L. Thonnard, Standardising outcome measurement in physical medicine and rehabilitation across Europe. Eur. Secur. **3-4**, 178–180 (1999)
320. S. Teraji, *The Cognitive Basis of Institutions: A Synthesis of Behavioral and Institutional Economics* (Academic, 2018)
321. L.L. Thurstone, *The Measurement of Values* (University of Chicago Press, Midway Reprint Series, 1959)

322. D. Torres Irribarra, R. Freund, W.P. Fisher Jr., M. Wilson, Metrological traceability in education: A practical online system for measuring and managing middle school mathematics instruction. J. Phys. Conf. Ser. **588**, 012042 (2015)
323. S.E. Toulmin, *The Philosophy of Science: An Introduction* (Hutchinson's University Library, 1953)
324. S.E. Toulmin, The construal of reality: Criticism in modern and postmodern science. Crit. Inq. **9**, 93–111 (1982)
325. C.A. Velozo, J. Lai, T. Mallinson, E. Hauselman, Maintaining instrument quality while reducing items: Application of Rasch analysis to a self-report of visual function. J. Outcome Meas. **4**(3), 667–680 (2000) http://jampress.org/JOM_V4N3.pdf
326. C.A. Velozo, Y. Wang, L. Lehman, J.-H. Wang, Utilizing Rasch measurement models to develop a computer adaptive self-report of walking, climbing, and running. Disabil. Rehabil. **30**(6), 458–467 (2008)
327. L.S. Vygotsky, Mind and society: The development of higher mental processes. Cambridge, Massachusetts: Harvard University Press (1978)
328. J.A. Weaver, A.M. Cogan, L. Davidson, T. Mallinson, Combining items from three federally-mandated assessments using Rasch measurement to reliably measure cognition across post-acute care settings. Arch. Phys. Med. Rehabil. **102**(1), 106–114 (2020)
329. T. Weitzel, *Economics of Standards in Information Networks* (Physica-Verlag, 2004)
330. L. White, *The Evolution of Culture* (McGraw-Hill, 1959)
331. A.N. Whitehead, *An introduction to mathematics* (Henry Holt and Co, 1911)
332. A.N. Whitehead, *Science and the Modern World* (Macmillan, 1925)
333. O.E. Williamson, The economics of organization: The transaction cost approach. Am. J. Sociol. **87**(3), 548–577 (1981)
334. G.L. Williamson, Exploring reading and mathematics growth through psychometric innovations applied to longitudinal data. Cogent Educ. **5**(1464424), 1–29 (2018)
335. M.R. Wilson, *Constructing Measures: An Item Response Modeling Approach* (Lawrence Erlbaum Associates, 2005)
336. M. Wilson, C. Carstensen, Assessment to improve learning in mathematics: The BEAR assessment system. J. Educ. Res. Dev. (Taiwan) **1**(3), 27–50 (2005)
337. M. Wilson, W.P. Fisher Jr., Psychological and social measurement: The career and contributions of Benjamin D. Wright, in *Springer series in measurement science and technology*, ed. by M. G. Cain, G. B. Rossi, J. Tesai, M. van Veghel, K.-Y. Jhang, (Springer, 2017) https://link.springer.com/book/10.1007/978-3-319-67304-2
338. M. Wilson, K. Scalise, Assessment of learning in digital networks, in *Assessment and Teaching of 21st Century Skills: Methods and Approach*, ed. by P. Griffin, E. Care, (Springer, Dordrecht, 2015), pp. 57–81
339. M. Wilson, K. Sloane, From principles to practice: An embedded assessment system. Appl. Meas. Educ. **13**(2), 181–208 (2000)
340. M.N. Wise, Precision: Agent of Unity and Product of Agreement. Part III – Today Precision Must Be Commonplace, in *The Values of Precision*, ed. by M. N. Wise, (Princeton University Press, 1995), pp. 352–361
341. L. Wittgenstein, *Philosophical Investigations (G. E. M. Anscombe, Trans.)*, 3rd edn. (Macmillan, 1958) (Original work published 1953)
342. F. Wolfe, D.M. van der Heijde, A. Larsen, Assessing radiographic status of rheumatoid arthritis: Introduction of a short erosion scale. J. Rheumatol. **27**(9), 2090–2099 (2000)
343. J.P. Womack, D.T. Jones, Beyond Toyota: How to root out waste and pursue perfection. Harv. Bus. Rev. **74**, 140–158 (1996)
344. A.W. Wong, S.F. Garcia, E.A. Hahn, P. Semik, J.S. Lai, S. Magasi, J. Hammel, K.P. Nitsch, A. Miskovic, A.W. Heinemann, Rasch analysis of social attitude barriers and facilitators to participation for individuals with disabilities. Arch. Phys. Med. Rehabil. **102**(4), 675–686 (2021)

345. A.W. Woolley, E. Fuchs, Collective intelligence in the organization of science. Organ. Sci. **22**(5), 1359–1367 (2011)
346. B.D. Wright, Sample-free test calibration and person measurement, in *Proceedings of the 1967 Invitational Conference on Testing Problems*, (Educational Testing Service, 1968), pp. 85–101
347. B.D. Wright, Solving measurement problems with the Rasch model. J. Educ. Meas. **14**(2), 97–116 (1977)
348. B.D. Wright, Foreword, Afterword, in *Probabilistic Models for Some Intelligence and Attainment Tests*, ed. by G. Rasch, (University of Chicago Press, 1980), pp. 185–199. [Reprint; Original Work Published in 1960 by the Danish Institute for Educational Research]. http://www.rasch.org/memo63.htm
349. B.D. Wright, Despair and hope for educational measurement. Contemp. Educ. Rev. **3**(1), 281–288 (1984)
350. B.D. Wright, Additivity in psychological measurement, in *Measurement and Personality Assessment*, ed. by E. Roskam, (Elsevier Science Ltd, 1985), pp. 101–112
351. B.D. Wright, Fundamental measurement for outcome evaluation. Phys. Med. Rehabil. State Art Rev. **11**(2), 261–288 (1997a)
352. B.D. Wright, A history of social science measurement. Educ. Meas. Issues Pract. **16**(4), 33–45 (1997b)
353. B.D. Wright, S.R. Bell, Item banks: What, why, how. J. Educ. Meas. **21**(4), 331–345 (1984)
354. B.D. Wright, J.M. Linacre, Observations are always ordinal; measurements, however, must be interval. Arch. Phys. Med. Rehabil. **70**(12), 857–867 (1989) http://www.rasch.org/memo44.htm
355. B.D. Wright, G.N. Masters, *Rating Scale Analysis* (MESA Press, 1982)
356. B.D. Wright, M.H. Stone, *Best Test Design* (MESA Press, 1979)
357. B.D. Wright, R.J. Mead, L.H. Ludlow, *KIDMAP: Person-by-Item Interaction Mapping (Tech. Rep. No. MESA Memorandum #29)* (MESA Press, Chicago, 1980) http://www.rasch.org/memo29.pdf. (6 p)

Author Index

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3

Subject Index

W. P. Fisher, Jr., S. J. Cano (eds.), *Person-Centered Outcome Metrology*,
Springer Series in Measurement Science and Technology,
https://doi.org/10.1007/978-3-031-07465-3