UNIVERSITY OF CALIFORNIA,
IRVINE

Statistical Methods in the Social Sciences

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Logic & Philosophy of Science

by

Kino Zhao

Dissertation Committee:
Professor Simon Huttegger, Chair
Associate Professor Cailin O'Connor
Chancellor's Professor Jeffrey A. Barrett

2021

DEDICATION

In Memory of

Kent Johnson

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my partner Greg Lauro, without whom I would have starved in my second year.

I owe immense thanks to my friends, who have kept me grounded, honest, supported, and intellectually challenged: Will Stafford for showing me the meaning of kindness; Mike Schneider for constantly igniting my love of philosophy; Jeff Schatz for always being there for me; Chris Mitsch for fighting along my side; Elliott Chen for making sure I never run out of milk tea or friendship; Jingyi Wu for the solidarity; and Kevin Kadowaki for always showing up and being dependable.

Additionally, I wish to thank Adam Chin, Kathleen Creel, Chrisy Xiyu Du, Sam Eklund, Margaret Farrell, Alysha Kassam, Jingjing Li, Chang Liu, Stella Moon, David Mwakima, Sarita Rosenstock, Hannah Rubin, Tim Schmitz, Darby Vickers, and Nicole Winter for accompanying me on this journey. I would like to thank and acknowledge Dr. Travis LaCroix, for being a friend but, more importantly, for sharing his LaTeX template with me.

I wish to thank a number of mentors who have believed in me: importantly, my dissertation advisor Simon Huttegger and committee members Cailin O'Connor and Jeff Barrett; my former mentors Sean Walsh and Kent Johnson, without whom I would have quit the program in my first year; and Holly Andersen and Nic Fillion, without whom I would not have gone into philosophy in the first place.

My work has benefitted greatly from invaluable discussions with scholars who who owe me no professional obligation but have listened to, encouraged, and critically engaged with my ideas: Konstantin Genin, Samuel Fletcher, Conor Mayo-Wilson, and Jan-Willem Romeijn. I thank Adrian Currie, Bill Harper, and Dan Hicks for treating me seriously and making me feel welcome in the profession.

I also wish to thank the administrative staff in LPS and the School of Social Sciences, especially Patty Jones and John Sommerhauser, for all of their hard, uncelebrated work.

Finally, I wish to thank my parents, for unconditionally supporting me even when they have no idea what it is that I do.

Chapter 1 is published in *Synthese*: 'Sample Representation in the Social Sciences' by Kino Zhao, 2020, reproduced by permission of Springer https://doi.org/10.1007/s11229-020-02621-3.

The entirety of this dissertation is conducted while in residence at University of California, Irvine, traditionally part of the homelands of the Acjachemen and Tongva peoples.

VITA

Kino Zhao

Ph.D in Logic & Philosophy of Science **2021**
University of California, Irvine *Irvine, California*

M.A. in Philosophy **2018**
University of California, Irvine *Irvine, California*

M.A. in Philosophy **2015**
Simon Fraser University *Burnaby, British Columbia*

B.A. in Philosophy, Psychology **2013**
University of British Columbia *Vancouver, British Columbia*

PUBLICATIONS

Sample Representation in the Social Sciences **2020**
Synthese

PRESENTATIONS

Sample Representation in the Social Sciences **2019**
Greater Cascadia History and Philosophy of Science workshop

A Statistical Learning Approach to a Problem of Induction **2018**
the 26th Biennial Meeting of the Philosophy of Science Association

Probabilism and Intuitionistic Logic **2018**
The University of Western Ontario's 18th annual Philosophy of Logic, Mathematics, and Physics Graduate Conference

A Model Theoretic Discussion of Statistical Learning **2017**
Women in Logic Workshop, Thirty-Second Annual ACM/IEEE Symposium on Logic in Computer Science

Statistical Learning and Reliable Processing **2017**
45th Annual Meeting of the Society for Exact Philosophy

Why Rationality? **2015**
University of Calgary - Graduate Student Conference

ABSTRACT OF THE DISSERTATION

Statistical Methods in the Social Sciences

By

Kino Zhao

Doctor of Philosophy in Logic & Philosophy of Science

University of California, Irvine, 2021

Professor Simon Huttegger, Chair

This dissertation is composed of projects on three aspects of gathering and learning from data in the social sciences: drawing representative samples, taking valid measurement, and making warranted inductive inferences.

Chapter one studies the challenge of drawing representative human samples. It is well documented that most samples used by studies in psychology-related fields are composed of Euro-American undergraduate students. Most writers agree that this is a serious problem for the generalizability of study results, but little improvement has occurred. By tracing the history of sampling, I identify the scientific and statistical rationale of sampling as a method of induction. I explain how the design-based approach, where to sample representatively is to sample randomly, became dominant. I show that this approach faces too many practical challenges within the social sciences to be useful as a guiding framework. In its place, I argue that the model-based approach, initially disfavoured for its theoretical shortcomings, is a better framework for the social sciences, because it allows the systematic integration of multiple imperfect samples. Instead of relying on one general framework to provide 'in-principle' justifications to all cases, the model-based approach allows context and background knowledge to inform practice.

Chapter two discusses measurement validity in the social sciences. Through an examination of the historical evolution of measurement and validity theories and the relationship between them, I argue that we should reject the view that measurement should be about an attribute that exists in the world in some robust sense, and that a pronouncement of measurement validity is a vindication of such an existence. First, I argue that this view, while attractive, has numerous theoretical difficulties and practical limitations. Next, I show that a rejection of this view, exemplified by the modern argument-based approach validity, presents a better perspetive in analyzing complex measurement problems in non-laboratory contexts. I conclude by pointing out that, based on the argument advanced in this paper, we should be more skeptical of ontological claims made on the basis of valid testing alone.

Chapter three studies the problem of induction in the context of statistical learning theory. I examine a claim in the literature that the Vapnik-Chervonenkis (VC) theorem, which specifies conditions under which a problem is machine-learnable, offers a response to the problem of induction. I prove that the problem of when this learnability condition applies in general is uncomputable. Hence this solution strategy fails. If statistical learning theory is trustworthy at all, the justification of this trust must be in parallel with other inductive methodologies and, consequently, subject to the same challenges.

I conclude this dissertation by arguing that a naturalistic, practice-first approach to philosophy of social science must pay attention not only to how a scientific method works in theory, but also to how it has been changed to accommodate resource-limited contexts.

Introduction

"Is there currently a crisis of confidence in psychological science reflecting an unprecedented level of doubt among practitioners about the reliability of research findings in the field? It would certainly appear that there is." This sentiment, expressed in Pashler and Wagenmakers (2012), marks the beginning of the meta-scientific worry that is now known as the *replication crisis*. From the apparent inability of the social psychology community to identify fraudulent research (Stroebe et al., 2012) to reports of negative replication results denied publication (Yong, 2012), discussions on Questionable Research Practices (Simmons et al., 2011; John et al., 2012; but also Fiedler and Schwarz, 2016) and statistical misuse (Simmons et al., 2011; McShane and Böckenholt, 2014) soon led to what Shrout and Rodgers (2018) called a "disciplinary panic". Scholars disagree not only on the solutions of the problem, but also on its nature, extent, or even existence, with some calling for radical intervention while others see it as a normal developmental phase of a young science (see, e.g., Pashler and Harris, 2012).

Laying at the heart of the mess is a series of previously underemphasized disagreements over how methods in the social sciences are supposed to work. If concluding significance at $p < 0.05$ is arbitrary, is there a nonarbitrary case of significance? What if every participant in the study shows the same effect? Does it matter *why* they agree or *who* they are? What, exactly, are we supposed to conclude when a replication study fails to reproduce? What are we supposed to conclude when a replication succeeds?

These questions are not foreign to philosophers of science. Since the very beginning, philosophers of science have tackled problems such as how to gather evidence, what evidence to gather, and how they support inference. Much of this discussion centers physics as the prototypical science, glossing over any differences between physics and the social sciences as signs of the immaturity of the latter.

Instead of forcefitting the social sciences into a philosophical frame molded from physics, pronouncing every misalignment as a sign of deficiency or “complexity” of the social world, this dissertation centers the social sciences. The chapters tackle three topics central to social-scientific methodology: sample representation, measurement validity, and inductive inference.

While these topics have been heavily debated in the context of the replication crisis in social psychology, my emphasis is not on the crisis itself. As mentioned before, the “unprecedented level of doubt”, whether or not it counts as a crisis, stems from the mismatch between expectations and reality, and yet there is remarkably little agreement over what these expectations ought to be. My main goal in this dissertation is to clarify these expectations in light of recent challenges they have faced.

The dissertation contains the following chapters. Chapter 1 studies the challenge of drawing representative human samples. It is well documented that most samples used by studies in psychology-related fields are composed of Euro-American undergraduate students. Most writers agree that this is a serious problem for the generalizability of study results, but little improvement has occurred. By tracing the history of sampling, I identify the scientific and statistical rationale of sampling as a method of induction. I explain how the design-based approach, where to sample representatively is to sample randomly, became dominant. I show that this approach faces too many practical challenges within the social sciences to be useful as a guiding framework. In its place, I argue that the model-based approach, initially disfavoured for its theoretical shortcomings, is a better framework for the social sciences, because it allows the systematic integration of multiple imperfect samples. Instead of relying on one general framework to provide 'in-principle' justifications to all cases, the model-based approach allows context and background knowledge to inform practice.

Chapter 2 discusses measurement validity in the social sciences. Through an examination of the historical evolution of measurement and validity theories and the relationship between them, I argue that we should reject the view that measurement should be about an attribute that exists in the world in some robust sense, and that a pronouncement of measurement validity is a vindication of such an existence. First, I argue that this view, while attractive, has numerous theoretical difficulties and practical limitations. Next, I show that a rejection of this view, embodied by the modern argument-based approach validity, presents a better perspetive in analyzing complex measurement problems in non-laboratory contexts. I conclude by pointing out that, based on the argument advanced in this paper, we should be more skeptical of ontological claims made on the basis of valid testing alone.

Chapter 3 studies the problem of induction in the context of statistical learning theory. I examine a claim in the literature that the Vapnik-Chervonenkis (VC) theorem, which specifies conditions under which a problem is machine-learnable, offers a response to the problem of induction. I prove that the problem of when this learnability condition applies in general is uncomputable. Hence this solution strategy fails. If statistical learning theory is trustworthy at all, the justification of this trust must be in parallel with other inductive methodologies and, consequently, subject to the same challenges.

Chapter 1

Sample Representation in the Social Sciences

1.1 Introduction

In 1936, the magazine *Literary Digest* set out to predict the US presidential election between Alfred Landon and Franklin D. Roosevelt. They surveyed more than 10 million people, of which 2.4 million responded, and concluded that Landon was going to win with 57% of the votes against Roosevelt's 43%. Instead, Roosevelt won with 62% against Landon's 38%.

This infamous incident is repeatedly cited to highlight the importance of selecting a sample that is representative. The *Literary Digest* employed a sampling procedure that favoured wealthy citizens over poor ones and did not correct for the vast majority of people who did not respond, resulting in a biased sample[1].

[1] The popular story told in statistics texbooks is that the *Literary Digest* used its own subscriber list, automobile registration and telephone books to choose its sample, and hence was biased towards wealthy Republicans (e.g., Likert, 1948; Scheaffer et al., 1971). This story is disputed by Bryson (1976), favouring instead the explanation from nonresponse bias.

The problem of sample nonrepresentation remains prevalent in the social sciences. For example, Sears (1986) analyzed the sample composition of research papers published during the year 1980 in three mainstream social psychology journals, *Journal of Personality and Social Psychology* (JPSP), *Personality and Social Psychology Bulletin* (PSPB), and the *Journal of Experimental Social Psychology* (JESP), and found the percentage of studies using American undergraduate students as samples to be 70% for JPSP and 81% for both PSPB and JESP. Their subsequent analysis of these journals for the year 1985 revealed no significant change. Arnett (2008) analyzed six prestigious psychology journals in different areas for the years 2003-2007 and found that most of the samples are taken from the United States (68%), with the remaining largely composed of people from other English-speaking countries (14%) and Europe (13%). A closer look at JPSP in 2007 reveals that 67% of American studies had samples consisted of undergraduate psychology students. More recently, Pollet and Saxton (2018) report that 79% of samples in the journals *Evolution & Human Behavior* and *Evolutionary Psychology* in the years 2015-2016 are from North America or Europe; moreover, 70% of the samples were either online samples or student samples. An analysis of three 2017 issues of *Psychological Science* by Rad et al. (2018) shows similar patterns.

The prototypical psychology sample, consisting of Euro-American undergraduate students, has been coined as WEIRD (Western, Educated, Industrialized, Rich, and Democratic) by Henrich et al. (2010b). Echoing researchers before them (e.g. Peterson, 2001, Wintre et al., 2001), they argue that we have considerable evidence to believe that the WEIRD subjects are very different from other people whom these subjects are often taken to represent.

While most agree that a WEIRD sample is not representative, articulating the exact desiderata of a representative sample proves difficult. In fact, some advocate abandoning the concept "representation" altogether, preferring the more concrete concept of random selection. In this paper, I argue that one major obstacle faced by the improvement of sample quality in the social sciences is the unrealistic expectation of a randomness-based conception of sample representation, called the design-based approach. Instead, I argue that a model-based approach to sampling offers a more realistic framework for effective assessment and improvement of imperfect samples.

This paper is organized as follows. Section 2 presents a brief history of how, through the seminal work of Jerzy Neyman (1934), random sampling superseded purposive sampling and became the preferred method among survey samplers. Representative samples, according to this framework, are ones generated through random selection. Section 3 points to difficulties with random sampling in practice and how, at least in the context of social science, falling short of the ideal standards result in systematic selection bias. Section 4 revisits the method of purposive sampling, re-emerging under the names of “model-based” or “prediction-based” approach through the works of Royall and Herson (1973). Based on this alternative approach, I argue for the old purposive sampling idea where a representative sample is one *balanced* on all relevant features. Section 5 discusses consequences of adopting a model-based perspective of sample representation in the social sciences and make practical proposals for improvement. Section 6 concludes.

1.2 Design-based Representation

The Norwegian statistician Anders N. Kiær is often credited as the first to bring sample-based research – that is, investigative methods which utilize only part of the population – to the attention of the western statistics community (Rao and Fuller, 2017; Smith, 1976; Kruskal and Mosteller, 1980). Around the turn of the 20th century, Kiær delivered a series of speeches at the annual meetings of the International Statistical Institute, advocating for the use of samples as effective proxies for studying populations.

The major source of skepticism Kiær faced was a lack of justification for sample-based inference, called the "representative method" at the time, which is inferentially ampliative. Kiær believed that we could identify a set of "rational selection procedures", produce "miniature populations", and draw accurate conclusions without full enumeration. He justified this approach empirically: he demonstrated that the sample-based survey results could be accurate but did not provide a theory for why the process worked (Seng, 1951). Other statisticians followed suit. Although few had comprehensive theories regarding why the representative method worked, many were able to demonstrate, empirically, that it did. Sampling was widely used in European government survey efforts by the 1920s.

With the increased use of the representative method in survey work, a new point of contention emerged between random and purposive selections. For our immediate purpose, the difference between them concerns whether the selection of a sample needs to be sensitive to the sample's composition. In random selection, the inclusion of each member into the sample is governed by probability alone, which is supposed to be identical across all members of the population. In purposive selection, the sampler aims to pick a fixed number of subjects with different characteristics so that the sample has the same proportions of those characteristics as the population.

We can see that these two sampling approaches correspond to Kiær's two conceptions of what a good sample should be. On the one hand, a representative sample should be drawn through "rational procedures", such as a random procedure that leaves no space for personal bias. On the other hand, a representative sample should be a "miniature population" in the sense of matching certain aspects of the population. The most natural way to achieve this goal is to deliberately select samples to be like the population in desired ways through purposive sampling. Although Kiær had both senses of representation in mind, they do not always coincide. That is, a sample drawn through a rational procedure may fail to be a miniature population. Consequently, samplers prefer one sense often had to let go of the other.

Although random and purposive sampling methods differ practically, they were not seen as direct competitors. Part of the reason may be that the dominant justification for the use of samples was still empirical – sampling with either method had been tried and true. In a 1926 Report for the International Statistical Institute, the English statistician Sir Authur Bowley distinguished the two sampling approaches and recommended them equally. All of these were changed by Jerzy Neyman's 1934 landmark paper.

Neyman's paper made two important contributions to the field of survey sampling. First, he provided a theoretical foundation for random sampling using his recently invented estimation method of confidence intervals. Second, he exposed an important flaw in purposive sampling. Although not everyone was convinced by Neyman's theory of confidence intervals[2], most were convinced enough to adopt random sampling as the superior method. Neyman's framework remained unchallenged within statistics until at least the 1960s. It is still very much the dominant paradigm among the social sciences today.

Neyman's definition of random sampling is elegantly summarized, in his own words, as follows (1934, p.585-586, emphasis original)

> Thus, if we are interested in a collective character X of a population π and use methods of sampling and of estimation, allowing us to ascribe to every possible sample, Σ, a confidence interval $X_1(\Sigma)$, $X_2(\Sigma)$ such that the frequency of errors in the statements
>
> $$X_1(\Sigma) \leq X \leq X_2(\Sigma)$$

[2] In particular, Bowley and Fisher remained skeptical, see Brewer et al. (2013).

> does not exceed the limit $1-\varepsilon$ prescribed in advance, *whatever the unknown properties of the population*, I should call the method of sampling representative and the method of estimation consistent.

There are two important claims of generality here. One of them is emphasized by Neyman, namely that the inference should hold regardless of the population distribution on the characteristic in question. This means that the success of the inference does not depend on assumptions made about the population. This generality resides in the heart of the supposed superiority of random sampling over purposive sampling. As will be discussed further in section 4, Neyman's primary criticism of purposive sampling is the fact that one would need to make a number of assumptions, many of which are either rarely true or rarely known to be true.

The other claim of generality is not highlighted or discussed much, which is the claim that the method of sampling should allow us to "ascribe to every possible sample" this desired property. In other words, it is the *sampling design*, rather than the sample itself, that justifies the inference. In fact, the justification of the inference should not refer to the specific characteristics of the sample at all. If an inference holds, it needs to hold for "every possible sample" drawn with the same method.

By putting the burden of justifying sample-based inference on the sampling method alone, to the explicit exclusion of referencing properties of the specific samples, Neyman's approach to sampling clearly follows the "rational procedures" line of Kiær's advocacy. Here, randomization is considered as the core of "rational" design, and it is in virtue of the power of design that the ampliative inference is justified.

Neyman's "design-based" approach to sampling remained unchallenged for decades. Later theorists developed more sophisticated schemes of sampling that allowed for uneven probability of inclusion across the population, but the basic idea remained. Randomization is the foundation of the representative method.

1.3 The Scientific Reality

According to the design-based framework, the inferential power of a sample comes from the sampling design, where the gold standard is random or probabilistic selection[3]. Theoretically, random sampling is often taken to contain two kinds of virtues. Smith (1983, p.394) explains,

> The arguments for randomization are twofold. The first, and most important for science, is that randomization eliminates personal choice and hence eliminates the possibility of subjective selection bias. The second is that the randomization distribution provides a basis for statistical inference.

I have explained that the statistical foundation of sampling is commonly considered to have begun with Neyman's 1934 paper[4], and yet sampling has been used widely before then. This is because random sampling, as a form of rational procedure, has a lot of intuitive appeal.

A central aspect of random sampling is the idea that the selection of elements is governed by probability, rather than scientists' intentions or other selection forces capable of causally influence the conclusions drawn from the sample. For example, suppose a group of surveyers is trying to estimate the average income of a country, then allowing the size of a person's

[3] I shall use the terms "random" and "probabilistic" interchangeably. Practically speaking, random selection implies that every element of the population has an equal chance of being included in the sample, whereas probabilistic selection allows that chance to differ from element to element. However, probabilistic sampling is almost always accompanied by a correction procedure where elements with greater chance of selection are weighed less in analysis. Theoretically, the two methods are the same.

[4] It seems that other statisticians, such as Bowley, have attempted to provide mathematical foundations for sampling before Neyman. However, Neyman does not discuss these alternative approaches in detail in his 1934 paper, and his paper is widely considered as the statistical landmark (see, e.g., Rao and Fuller, 2017 and Srivastava, 2016). It seems reasonable to conclude that whatever mathematical foundations of survey sampling existed before Neyman, whether or not they are adequate, have had limited historical influence.

house to influence the probability of that person being included in the sample is going to result in biased estimations. To guard against a tendency to preferentially sample people with big hourses or small ones, one needs to make sure that the size of someone's house cannot inform the probability of them being selected into the sample. The best way to achieve this goal regarding not only house size but all other forms of influence is to make the selection procedure maximally uninformative. Random selection is, at its core, a maximally uninformative selection procedure.

This intuitive appeal of random selection relies on the premise that maximal noninformation is sufficient in removing undesired interference to study results. As the phrase suggests itself, maximal noninformation precludes outside factors from systematically affecting ("informing") a sample's composition. However, this does not mean that the undesired biases would not occur.

To better appreciate this worry, consider again the problem of sample nonrepresentation discussed in section 1. In their influential attack on WEIRD samples, Henrich et al. (2010b) specifically argued that the worry with WEIRD samples is not simply that most of the world's population is not WEIRD, but that the behaviours of WEIRD samples may differ substantively from the rest of the population. They specifically sited results from Segall et al. (1966)[5] on how people from some cultures are not subject to the Müller-Lyer illusion and from Henrich et al. (2010a) on how people from different cultures respond to the Dictatorship and Ultimatum Games differently as reasons to be skeptical of the generalizability of results obtained from WEIRD samples. The point of contention from their critics is also that many behaviours are not subject to cultural influence. For example, Gächter (2010) argues that whether the use of student samples is problematic depends on the research question. In

[5] A reviewer has pointed out that the validity and interpretation of Segall's results have been disputed. Indeed, it is a persistent difficulty to determine whether an observed difference is due to a difference in sample composition, methodological variation, or a number of other factors deemed irrelevant. One goal of the framework advocated in this paper is to help better systematize the variations in sample composition so as to facilitate better hypothesis testing regarding the source of a variation.

particular, since economic behaviours are taken to be universal, “any subject pool is in principle informative about whether theoretical predictions or assumptions contain behavioral validity” (p.2).

It is clear that the problem with biased samples is not so much that members of the sample “look” very different from members of the population. Instead, the worry is that these apparent differences translate to unappreciated behavioural differences and so the results obtained from the sample are not generalizable to the population. Similarly, the worry with “subjective selection bias” is not so much that a bias results in members of a sample more likely to have certain characteristics, but rather that *these characteristics interfere with drawing accurate conclusions from the sample.*

This form of systematic bias often occurs as a result of “personal choice” in the sense of preferential sampling, which may happen consciously or unconsciously. A researcher may consciously choose to sample wealthier citizens as a way to inflate the national average income estimation. Alternatively, the researcher may unconsciously choose to sample only those who are dressed nicely, leading to the same effect. From the perspective of drawing conclusions from a sample, both forms of personal choice result in undesired systematic bias. Random selection eliminates both sources of influence.

However, the same bias may also occur as a matter of chance. Even if the sampling procedure is truly random, it is still possible that a particular sample happens to consist of members who are wealthier than the national average. To see why this is the case, consider how, even though a fair coin has a 50% chance of landing head, it is not the case that, for every 10 coins I flip, exactly 5 of them will land heads. If the coin is truly fair, the Law of Large Numbers guarantees that, as the number of flips goes to infinity, the proportion of heads converges to the true proportion – 50%. However, the Law of Large Numbers does not guarantee that the true proportion will be reached at any finite stage. In fact, it does not even guarantee that my estimation always improves with more flips[6]. Similarly, random selection only guarantees that, if the population is repeatedly sampled for infinitely many times, then the average of the sample means approximates the population mean. It does not guarantee that any single sample will have the same mean as the population.

The procedure standardly used to address the problem of chance bias is *post-stratification*. In stratified sampling, a population is divided into multiple mutually exclusive, collectively exhaustive "strata". Samples of different sizes are drawn, randomly or otherwise, from these strata. The resulting samples are weighed by the size of their strata relative to the population and combined to form the final sample. In post-stratification, the process is reversed. After a sample is drawn, it is partitioned into groups, often along some salient characteristics deemed important by the researchers. The associated strata are reversely constructed, their relative ratio computed from auxiliary full population data, and the groups weighed accordingly.

Consider the National Comorbidity Survey (NCS) as an example, which was launched in the US as "the first psychiatric epidemiologic survey to administer a broadbased research diagnostic interview to a nationally representative sample of the United States" (Kessler, 1994). The NCS uses a stratified, multistage area probability sample, which is common for survey efforts of its scale. The core sample contained 47.5% males. However, according to the National Health Interview Survey (NHIS) of 1989, a full enumeration of the population rather than a sampled survey, 49.1% of Americans were male. The NCS therefore post-stratified their sample by giving more weight to results obtained from the sampled males than that of the females.

[6] In certain special cases and with strong additional assumptions, a method may guarantee uniform convergence, where the estimation is always improved with increased sample size. When that happens, one can obtain an ε-δ bound on how far "off" we can be for a given confidence threshold. However, this option is only open for fields where it is easy to repeatedly, truly-randomly gather large samples, which is unrealistic for the social sciences.

It is worth noting that the NCS, despite adopting a method as close to random sampling as is feasible, still feels the need to adjust data to compensate for sample imbalance. This shows the limitations of random selection as a guard against chance bias.

More significantly, the method of post-stratification does not really fit in the design-based framework. Recall that the design-based conception of sample representation relies exclusively on the power of the selection process to justify sample-to-population inference. The idea is supposed to be that, as long as researchers adopt adequate sampling procedures, they should not feel the need to also analyze sample composition.

Besides, the design-based framework does not provide guidance for how sample composition should be analyzed. To see this, we can compare the NCS with similar sampling efforts from other countries. The NCS of America post-stratified against sex, age, marital status, race, education, region, and urbanicity (Mickelson et al., 1997, p.1095); the German National Health (GNH) survey post-stratified against sex, age (with a different range), marital status (in finer categories), and employment status (Jacobi et al., 2002); the Australia National Mental Health Survey (ANMHS), however, decided to not post-stratify at all (Henderson et al., 2000).

In addition to the inconsistencies across similar survey efforts, those that do post-stratify provide very little reasoning as to why they decide on the characteristics that they do. Post-stratification as a method depends on the existence of full enumeration demographics data like the NHIS, which is often called "auxiliary data" or "organic data" in this context. The existence of such data limits whether and how a sample survey can afford to post-stratify. That said, post-stratification also reflects conscious choices on the part of the research team. For example, the NCS chose to post-stratify against the NHIS rather than the US Census because "[the NHIS] includes a much wider array of sociodemographic variables for the purposes of poststratification" (Mickelson et al., 1997, p.1095). This is certainly not because the US Census did not gather a lot of data. In 1989, the Census Bureau gathered information as diverse as age differences between bride and groom, prevalence of AIDS, immigrative status, and average weekly expenditure (US Census Bureau, 1989). Instead, the US Census gathered *the wrong sorts of data*, at least from the perspective of the NCS.

It is clear that researchers make judgments about which characteristic imbalance is worth correcting in a randomly selected sample, and yet these judgments are rarely explicitly stated or argued for. Indeed, there is no theoretical space within the design-based framework for such corrections, so it only makes sense that corrections like these, when they do occur, are guided more by intuition than by arguments.

The discussion concerning post-stratification has highlighted two important observations. First, even the best random selection efforts result in sample imbalances deemed worthy of correction by researchers. The elimination of “subjective selection bias” guaranteed by random selection is clearly insufficient. Second, while post-stratification is frequently used to correct for chance bias, the practice is not principled. This is because post-sampling corrections of this form do not fit into the design-based understanding of how sampling is supposed to work.

Worse still, large-scale survey efforst like the NCS are relatively uncommon; most research teams within the social sciences do not have nearly as much resources to employ anything like an area probability sample over a nation. This is compounded by the fact that many research projects within psychology, anthropology, and economics target the entire humanity as the intended population. If random sampling over a country is difficult, random sampling over the entire human race is practically impossible. This is especially true when the study procedures are very involved, such as in experiments, longitudinal studies, or when data are collected qualitatively.

Another major obstacle for samplers in the social sciences is the problem of nonresponse. When the sampled units are humans, there is always a chance that someone sampled will decline to participate. When that happens, the actual sample will differ from the theoretical sample envisionsed by design. Since nonresponse makes a probabilistically selected sample effectively nonprobabilistic, it is a serious problem. Indeed, many believe that failure to address nonresponse is the true culprit behind the epic failure of the *Literary Digest* poll.

The coping strategy developed in the 1950s, which continues to be the preferred strategy today, is two-phase sampling. In the first phase, the preferred measurement procedure is used for everyone theoretically selected in a sample. If some members of the sample do not respond, then a second phase is carried out where a different measurement procedure is used to reach nonrespondents. The idea is that the alternative measurement, while less ideal in other ways, may change the minds of nonrespondents. For example, the alternative method may be a more resource-intensive in-person interview as opposed to a paper-based questionnaire, or it may be a shortened version of the questionnaire which takes less time for subjects to complete. If the second phase elicits near-full response and the alternative measurement methods are considered empirically equivalent, then the problem of nonresponse is fully corrected. If significant nonresponse remains at the second phase, researchers would often assume homogeneity among nonrespondents and post-stratify as if second phase nonresponse is undersampling. Unsurprisingly, nonresponse remains a serious challenge today.

When the sampling procedure cannot plausibly be construed as random, the design-based framework ceases to provide guidance. Statistical foundations for the design-based approach, such as Central Limit Theorems, rely on the conceptualizaion of sampled elements as random variables. From this perspective, all non-random selection procedures are equally bad.

What this also means is that, if a study cannot obtain random selection, researchers lose any sense of how they might still improve their sample. The essence of probabilistic sampling is that every member of the population has a non-zero probability of being selected. Even if this probability varies from member to member, post-sampling correction methods such as post-stratification can adjust the weights such that the results are “as if” selection is truly random.

However, if some members of the population have probability 0 of being selected, then there is nothing one can do to make the data look as if those members could have contributed. One cannot modify nonprobabilistic samples *post hoc* to make them probabilistic. I believe this is the main reason that, despite wide recognition of the problem of sample nonrepresentation, the proportion of studies employing undergraduate-only samples has not changed over the decades.

In the absence of principled ways of improvements, convenience becomes a major driving factor. Convenience sampling refers to the practice where members of the sample are chosen because of ease of access and recruitment. The most common form of convenience sampling is using undergraduate students at the same institution where the researchers are based. Other forms include Amazon Mechanical Turk or community members recruited using posters or email advertisements.

Unsurprisingly, convenience sampling is the most common form of sampling within the social sciences. An analysis of sample composition in 5 journals in developmental science shows that 78-88% of all studies published in years 2007-2011 that use American samples use convenience sampling (Bornstein et al., 2013). Given the prevelance of undergraduate and online samples within the social sciences, the same is likely true of other fields as well.

In addition to being nonprobabilistic, convenience sampling often perpetuate a specific kind of systematic bias. Consider, again, the use of WEIRD samples, where E stands for educated and R for rich. It should not be surprising that people who are rich and educated are more likely to have the leisure to participate in odd psychological studies. This is especially true when the study offers very little compensation, as is the case in most resource-limited academic contexts.

The prevalence of convenience sampling highlights an important feature of sample design that is often overlooked in abstract discussions – sampling involves not only a decision about design, but also a series of actions associated with actually contacting and recruiting subjects. Without an explicit intention to guard against this tendency, subjects who are more "accessible" are likely going to dominate conveniently gathered samples. This is especially problematic because subjects who are less accessible are usually such because of other forms of marginalization. For example, one limitation identified by the ANMHS is the noninclusion of indigenous people who live in remote locations (Andrews et al., 2001). For another example, the persistent underrepresentation of African Americans in samples used in clinical psychology studies (Graham, 1992) is likely to be a major contributor to the persistent clinical malpractice disproportionatelly experienced by this population (Hall, 1997).

To summarize, the design-based framework of sample representation, where random sampling is considered the gold standard, faces two major problems in practice. First, while faithful execution of random sampling can eliminate intentional selection bias, it cannot eliminate chance bias. The scientific importance of chance bias can be witnessed by the wide use of post-stratification as a correction mechanism. However, the design-based framework provides no guidance for such corrections, which is why they are often carried out inconsistently and with little justification. Second, random selection is extremely difficult to achieve in resource-limited contexts. When random selection is out of the question, the design-based framework is again silent on how a sample may still be improved. Consequently, researchers rely on convenience as the dictating principle. Convenience sampling often introduces systematic bias of a particular kind that are likely to compound existing social gaps.

1.4 Balanced sampling

To briefly return to the history of sampling, recall that sample representation was used in two distinct senses by Kiær and his immediate followers: as samples obtained through "rational selection procedures" and as samples that are "miniature populations". Since these senses do not always coincide, differing priorities have led survey samplers to two different paths: random sampling and purposive selection. Neyman's 1934 seminal paper convinced the statistical community that random sampling rests on a solid statistical foundation, while purposive selection relies on contentious and unrealistic assumptions.

In purposive sampling, one has a variable of interest, X, and a number of control variables. For ease of illustration, assume there is only one control variable, Y. In the early form of purposive sampling targeted by Neyman, X and Y are assumed to be linearly correlated. Assume the characteristic Y is well known and easily measured, one can purposively select members such that the sample distribution on Y matches that of the population[7]. This sample is considered representative with respect to X.

Neyman's criticism of purposive sampling consists of two aspects. First, he pointed out that the then-recent Italian sample survey, which used purposive selection, was vastly inaccurate – this form of empirical argument has always carried a lot of weight with surveyers. Second, Neyman pointed out that the assumption of linear dependence between X and Y is often unrealistic. Random sampling, Neyman explains, is assumption-free.

As statisticians delved deeper into the foundation of sample-based estimation, they gradually reazlied that random sampling, or at least inferences based on it, are not as straightforward as Neyman had believed. For example, Godambe (1955) developed a unified account of a class of estimators commonly used around that time and showed that there does not exist a best linear unbiased estimator in this general class, contrary to what Neyman had claimed. Later, he showed how the likelihood function from the full sample data, theoretically understood to include a set of labels together with associated variables of interest, provides no information

[7] Neyman's original analysis was based on stratified versions of random and purposive sampling. In his rendition of purposive selection, each stratum was sampled such that the mean of Y in the stratum sample equaled the mean of Y in the overall stratum. Allowing the means of Y to differ among strata, Neyman's description of stratified purposive sampling is equivalent to sampling from the entire population in a way that the sample distribution of Y matches the population distribution of Y.

on the non-sampled values and hence on the population total or mean (Godambe, 1966; see also Rao and Fuller, 2017).

Against the backdrop of theoretical and practical challenges to random sampling, a new approach was developed by, primarily, the statistician Richard Royall (Royall 1968, 1970, 1992; Royall and Herson, 1973). Royall's basic observation is that sample-based inference can be conceptualized as a prediction problem, where results obtained from sampled individuals are used to predict what results we would obtain from the unsampled part of the population.

According to the design-based framework, the inferential power of the sample comes from the idea that, while *these* individuals were in fact sampled, the sample could very easily have contained *those* other individuals instead – those ones we are trying to estimate. In other words, the members in the sample are *interchangeable* with members outside of the sample in some sense[8]. From a prediction perspective, however, the relationship between sampled and unsampled individuals need not be nearly as strong. If I am using a person's wealth to predict their life expectancy, I do not need to assume that the tax return data, say, I have obtained from one person could have been from another person instead. What I do need to assume is that the person whose data I have is sufficiently similar in relevant ways to the person whom I'm trying to predict.

If the two tax returns are considered two instantiations of the same random variable, as in the case of design-based random sampling, then the assumption that they should be similar is in some sense warranted. However, if we know what it means for two people to be similar in this context, then we can check whether they indeed are similar in an *ad hoc* way. For example, if we think a person's country of residence affects their life expectancy, then we

[8] Exchangeability is a Bayesian perspective on how random sampling works. The design-based framework is, by and large, developed and used under the frequentist paradigm, where random selection is defined as i.i.d. (independent and identically distributed) sampling, which grounds the application of the Law of Large Numbers. Exchangeability is presented here because it offers are more intuitive description of the inferential process.

would want to make sure that the sampled person resides in the same, or a relevantly similar, country as the unsampled one.

This way of understanding sample-based prediction leads to the form of purposive sampling targetted by Neyman – if we believe that matching distribution of Y between sample and nonsample ensures that the individuals from these two groups are similar, then we should sample to match distribution of Y. If Y is indeed the only characteristic correlated to X, then the resulting sample would be a "miniature population" in Kiær's sense – it mirrors the population in a way relevant to the study target, X.

A major contribution by Royall is to develop a more general account of this style of inference where a sample that does not already match the population on the entire distribution of Y can be used in similar ways with extra assumptions. For example, in ratio estimation, the ratio between X and Y is considered constant along different values of Y. Suppose a person's wealth and their life expectancy are positively linearly correlated and that the sample we have consists mostly of people wealthier than the national average. In this case, we can compute the slope of the trendline relating X and Y from our sample of wealthy subjects and extrapolate this information for poorer ones. If we know the national average wealth, we can estimate the national average life expectancy accordingly. All of this is done without referencing the sample gathering process.

As is evident from the above example, this approach, called the model-based or prediction-based approach, relies on a number of auxiliary information. We need to first identify one or some control variable(s) Y, with the assumption that they relate to X strongly enough to serve their intended function. We also need certain kinds of population-level data concerning Y. If we cannot match Y across the entire distribution – which is almost always true in practice – we will need to make assumptions concerning the nature of the relationship between X and Y. In the case of ratio estimation, the model requires a linear dependence between X and Y that passes through the origin. With increased computational power, more complicated dependence relationships can be accommodated.

With as many assumptions as needed in even the simpliest cases, the problem identified by Neyman is a serious one. Just as we may be wrong about the relationship between X and Y being linear, we may also be wrong about any other assumed nature of this relationship, or that they are related at all. This problem is one of model misspecification, which is always a challenge in model-based inference. Indeed, model misspecification, especially the kind that is difficult to detect but can significantly bias the resulting estimation, has been the major challenge to the model-based approach (Hansen et al., 1983)[9].

Royall and Herson (1973) showed that sample balance can protect against model misspecification. Their definition of sample balance is as follows. Suppose $Y_1 \ldots Y_n$ are all the variables upon which X is dependent. Then a balanced sample is one where the mean of each Y_i $(1 \leq i \leq n)$ of the sample equals that of the population. If a sample is balanced in this way, then many model-based estimators retain their optimality and unbiasedness under many instances of model misspecification[10].

Model misspecification remains a threat as long as sample balancing is practically difficult. In the social sciences, researchers often choose to study X precisely because they do not know how it relates to other variables. Questions of model misspecification and sample balance are intrinsically part of the unknown. In other words, if researchers knew that the model was adequate or that the sample was balanced, they would not have conducted the study in the first place.

[9] A design-model hybrid approach, called model-assisted sampling was developed not long after the development of the model-based approach. The hybrid approach aims to use properties of random selection to help guard against model misspecification (Cassel et al., 1976; see also Brewer, 1999). I will not discuss the hybrid approach for two reasons. First, the importance of purposive balancing, which is my main thesis, is equally emphasized in both the model-based and hybrid approaches. Second, the guarding power of the hybrid approach against model misspecification only appears in large samples with relatively good randomization, which is not part of my target.

[10] These estimators are approximately unbiased if the sample is approximately balanced.

The immediate consequence of this observation is that, like random sampling, the model-based approach does not offer an easy route to sample representation in the context of resource-limited social sciences. This should not come as surprise, however, as a change in perspective is not supposed to magically solve an intrinsically difficult problem. The benefit of the model-based approach is that it provides a framework capable of guiding sample improvement in systematic ways.

Recall that one important shortcoming of the design-based framework discussed in the previous section is that, once a research team cannot obtain probabilistic sampling, it is difficult to see any other ways of improvement. In the absence of such principled guidance, convenience becomes the dominant consideration, leading to systematic bias. The benefit of the model-based framework is that one can explicitly state all the assumptions necessary to support the inference in question and discuss the evidence we have of them.

Return to the example of a person's wealth and life expectancy. In order to propose that our ratio estimator based on a biased sample of wealthy individuals is adequate in estimating the national average life expectancy, we need to make the following assumptions. First, variation in wealth accounts for most of the variation in life expectancy. Second, the relationship between wealth and life expectancy is positively linear, with the regression line passing through the origin. Third, our sample of wealthy individuals, albeit biased, contains enough data points to accurately estimate the slope of the regression line. Fourth, our information regarding the national average personal wealth is accurate.

Once explicitly laid out, skeptical researchers can challenge these assumptions methodically. For example, some may argue that wealth has only limited influence on life expectancy, or that the influence is moderated by the nature of the person's job. Others may argue that the contribution of wealth to life expectancy has a diminishing marginal return, where the increase in wealth produces less impact for wealthier individuals than for poorer ones. On the one hand, each of these challenges cast doubt on our claim that the ratio estimator is adequate. On the other hand, each of these doubts can be addressed with auxiliary evidence.

The same thought process can be used for preemptive improvements of samples, too. For example, I may believe that, in addition to wealth, the number of children a person has also predicts their life expectancy. If such information is not difficult to obtain, I may decide to have my sample of wealthy individuals balanced on the number of children they have even if I do not have perfect evidence concerning the nature and extent of the effect of children, with the knowledge that this additional act of balance is always beneficial. Furthermore, I can even perform a kind of cost-benefit analysis between convenience and theoretical improvements. Suppose, for example, that I have reasons to believe that the relative importance between job type and number of children to a person's life expectancy is comparable, and yet information regarding job type is much more difficult to obtain. I may choose to balance my sample against the number of children but not job type. This will make my inference less than perfect, but still better than using wealth alone.

This style of thinking is already present in the use of post-stratification, if only implicitly. Recall that post-stratification is a method aimed at balancing an already-drawn sample along some selected characteristics. The method is widely used in the design-based setting, but receives no theoretical guidance from the framework. Consequently, the choice of which characteristics to post-stratify against tends to be inconsistent across similar survey efforts. From the model-based perspective, however, post-stratification makes perfect sense. Indeed, ratio estimation from a biased sample can be seen as a form of post-stratification (Smith, 1991).

From the model-based perspective, researchers should post-stratify against characterstics they believe to be statistically relevant to the target variable. Limited by the availability of auxiliary data, researchers may choose to post-stratify against only variables they believe to contribute significantly or feel that they have strong enough evidence for believing so.

Differences in such subjective thresholds can lead to inconsistencies in post-stratification decisions across similar survey efforts, as observed.

To summarize, model-based inference in sampling relies on assumptions concerning the relationship between control and target variables. To guard against possible inaccuracies in these assumptions, a sample should be balanced, either through purposive design or post-stratification. An ideally balanced sample – one that is representative in the "miniature population" sense – guards against many forms of model misspecifications, whereas an approximately balanced sample approximately guards against model misspecifications. The adequacy of the model-based framework is attested by its ability to account for both the intuitive justification and the practical inconsistencies observed with post-stratification.

In extremely resource-limited cases, even an approximate balance may not be feasible. Suppose I am using a sample to study how much people, in general, are willing to share their newly acquired wealth with a stranger. I may have some suspicions, evidentially justified or not, that certain characteristics could affect the extend of giving in a systematic way. For example, perhaps those who have gone through financial hardships themselves are more likely to empathize and share with strangers. Note that these suspicions appear a lot like independent variables in an experiment – indeed, one could systematically study altruistic behavioural differences between the rich and the poor. However, even when between-group differences are not of theoretical interest, cross-group balance is still important from the perspective of sample representation.

Nevertheless, I may not have a good understanding of how levels of wealth affect altruistic behaviour or a feasible way of balancing my sample across levels of wealth distribution. Moreover, it is highly likely that, even if wealth plays a role, its relationship with altruism accounts for only a small proportion of the total variation, and I may not have any idea at all what other variables are worth controlling for.

Situations like these are common in the social sciences. While balancing the sample against one more variable takes us closer to the ideal of full representation, controlling variables that only account for a minority of the total target variance is not sufficient to eliminate systematic bias. However, the benefit of the model-based framework is not about meeting the same standard with less content, but rather documenting and systematizing available and unavailabe information in a way that makes assessment possible.

Suppose I am able to secure participants at the top and bottom levels of the wealth hierarchy. This act of balancing accounts for wealth if wealth is linearly related to altruism, but not if they are quadratically related. If later research finds the relationship to be quadratic, then others can reasonably question the accuracy of my results. Similarly, if later research finds that age, a variable I have not controlled for, is also statistically related to altruism, then that would similarly constitute a weakness of my initial estimation.

More importantly, reasoning like above allows for better synthesis of similarly aimed research. Suppose I conduct a study on altruism with a sample controlling for wealth only, and another research team conducts a similar study controlling for age only, and that our results are very similar. The model-based framework allows us to infer that, if someone had drawn a sample balanced on both wealth and age, they would have also gotten similar results. While multiple nonprobabilistic samples cannot be combined to form a probabilistic sample, multiple samples balanced in different ways can be combined to form a sample that is balanced on all of those ways. The model-based framework, therefore, allows for a systematic integration of resource-limited studies.

1.5 Sample representation in the social sciences

In an attempt to address the problem of questionable research practices in psychology, Simons et al. (2017) proposed that research papers should be required to include "Constraints on Generality" (COG) statements in their methods sections. They describe their vision as follows (p. 1124),

> A COG statement specifies your intended target population and the basis for believing that your sample is representative of it; it justifies why the subjects, materials, and procedures described in the method section are representative of broader populations.

Focusing on the sampling aspect alone, the proposal provides little guidance asside from *justifying why the sample is representative.* In this paper, I have discussed two senses of sample representation: the design-based approach where a representative sample is one drawn randomly and the model-based approach where a representative sample is balanced on all features relevant to the research target. I have further argued that the ideal versions of both senses of representation is infeasible for most research groups. Demanding a research team to explain why their sample, gathered with severe resource limitations, is representative is unlikely to lead to tangible improvements. We need proposals that are more feasible for small-scale research efforts.

Despite the dominence of the design-based framework in the social sciences, the idea that a representative sample is one that is balanced on relevant features is not foreign. Studies that use samples often report participants' demographics, which is how the literature was able to detect the lack of sample representation in the first place. However, few, if any, document reasons for why they choose to report the type of demographics that they do. As in the case of post-stratification, it seems reasonable to suppose that researchers are making these decisions based on the intuition that a more balanced sample is a better sample. If we take a model-based perspective, we can begin to unpack these implicit assumptions and question their adequacy.

Because of the dominance of the design-base framework, most metascientific studies on sample representation focus on sampling method rather than sample composition[11]. Nevertheless, a few studies have examined the practice of demographics reporting. In an analysis of all studies published in four pediatric psychology journals in 1997, Sifers et al. (2002) reported that “participants’ ages, genders, and ethnicity were reported at moderate to high rates, whereas socioeconomic status was reported less often”. Of the 260 papers they analyzed, gender was reported in 86.2% of the papers, whereas SES was reported in 46.5%. A similar review of studies published in *Psychological Science* in 2014 found that, while gender is reported in 75% of the studies, education levels is reported in only 52%, race/ethnicity is reported in 20%, and SES in only 8% (Rad et al., 2018).

Reporting sample demographics, even without explicit efforts at balancing or post-stratification, allows later researchers to better assess the overall coverage of the literature. That said, asking researchers to report “as many demographics as possible” is also infeasible. A lengthy demographics questionnaire attached to all studies is likely to cause cognitive fatigue in participants, harming study validity. There are also privacy concerns over potential reidentification through aggregated demographics data.

In other words, the control over sample demographics, be it actual balancing or mere reporting, requires deliberate planning. This is especially true in studies with small samples that are all gathered from the same location and through the same method, both of which in-

[11] Although the acronym “WEIRD” refers to a set of demographic features, the metascientific data Henrich et al. (2010b) relied on primarily concerned *where* samples were drawn, e.g., from undergraduate psychology classes at the researchers’ universities, supplemented by secondary data on the demographics of students of such universities.

strinsically limit the diversity of the sample. Consequently, researchers need to be deliberate in choosing which control variables to report.

According to the model-based framework, an estimation is unbiased just in case all of the *statistically relevant* variables have been controlled for. This means what variables are worth controlling will change depending on what the research target is[12]. Instead of asking researchers to always report as many control variables as possible, it is more effective to report only a few that are considered statistically relevant to the target and explicitly justify them as such.

Furthermore, to assess the balance of a sample, researchers need auxiliary data concerning population-level composition. Balance is important for any study aimed at sample-to-population generalization, even when demographics is not part of the research interest. Consequently, researchers of human subjects in any discipline should pay attention to how individual characteristics systematically affect behaviour, as well as how such characteristics are measured in full enumeration survey efforts. Sociologists are beginning to notice the mismatch between the changing societal understanding of sex and gender and the traditional ways of measuring them (Westbrook and Saperstein, 2015; Hart et al., 2019). Similar forms of close scrutiny of the methodology and assumptions underlying demographic surveys should become a bigger part of all areas of the social sciences, not just demography.

[12] Interpreted from this perspective, the preferential reporting of gender as a control variable brings up a series of questions concerning the presumed roles (and the presumed univocality of such roles) gender plays in shaping behaviour. Similar observations can also be made about the overreporting of some demographic variables and the underreporting of others. Indeed, since design-based principles cannot guide reporting or poststratification, culturally entrenched ideologies often substitute for this role. The philosophical implications of this dynamic are beyond the scope of the current paper but will be the subject of future work.

1.6 Conclusion

The social sciences face a persistent problem of sample nonrepresentation with no trend for improvement. I believe this is due to a lack of feasible proposals for resource-limited contexts. By tracing the history of sampling, I showed how the design-based framework for sampling where random selection is the gold standard, although good on paper, provides little practical guidance when the gold standard cannot be achieved. In contrast, the model-based framework provides a systematization of all assumptions, allowing them to be challenged and defended methodically. It also offers guidance on how small-scale studies with imperfect samples can be integrated for greater understanding.

Accordingly, I have made two practical proposals for the improvement of sample representation in the social sciences. First, instead of inconsistently reporting a more-or-less identical set of sample demographics, researchers should deliberately select a few that they believe to be statistically relevant to their research target and explicitly justify them as such. Second, there should be greater communication between scientists studying human behaviour and demographers designing full enumeration survey efforts.

Chapter 2

Measuring the Non-Existent: Validity Before Measurement

According to the latest edition of the *Standards for Educational and Psychological Testing* (2014), published collaboratively by major psychology and education associations and representing the “the gold standard in guidance on testing in the United States and in many other countries” (American Psychological Association website), validity is “the most fundamental consideration in developing tests and evaluating tests” (*Standards*, p.11). Yet the need for a theory of validity distinct from any theory of measurement should strike a naive reader as odd: what is a theory of how to measure if not also a theory of how to measure well?

Motivated by this question, the present paper explores the changing conception of validity, its relationship with theories of measurement, and what this dynamic means for ontology. To give a brief preview, validity played no role throughout the development of the Representational Theory of Measurement and meant little more than generic endorsement of test quality in early psychometric theories. The first major shift came with Cronbach and Meehl’s (1955) influental paper on construct validity, published as a kind of philosophical supplement to the 1954 *Technical recommendations for psychological tests and diagnostic techniques*, the precurser of the *Standards*. The construct validity program attempted to provide a *metaphysical* foundation for theories of measurement – at least as metaphysical as is allowed by the logical empiricist framework, from which this foundation derives. Later I will show how both the perceived need of a metaphysical foundation of measurement and the robust theorizing that followed were responses to a set of epistemic problems ubiquitous to all forms of measurement. The construct validity program eventually failed, partly due to the decreased popularity of logical empiricism, but also partly due to practical difficulties associated with the broadened use of tests. Its replacement, sometimes called the *argument-based approach* to validation, insists that validity is not about measurement at all, but what we do with measurement results, a position still disputed today.

Along this historical narrative, I will argue that validity theory's shift away from measurement represents a radical reconceptualization of the metaphysics of measurement, one that allows us to know what it means to measure well *before* we know what it means to measure. While this picture may sound counterintuitive, it stands in better coherence with other contemporary philosophical theories of measurement, particularly Hasok Chang's re-analysis of operationalism and Theodore Porter's theory of quantification as technology.

Insofar as the validity-before-measurement picture is possible, a natural question will arise concerning its scope of application. One temptation, which I will resist, is to draw a line between the social and the physical sciences: the physical sciences do not need validity theory; the social sciences do. However, rejecting this crude dichotomy is not part of my main thesis. Instead, my emphasis will be on what the validity-before-measurement picture tells us about the nature of measurement-based evidence. In particular, I will argue that the possibility of having validity theories without measurement theories shows that having valid measurement is insufficient in supporting a kind of ontological claim about the world that is sometimes made on grounds of valid measurement alone.

The paper is organized as follows. In section one, I review pre-1955 measurement theories, which were early forms of the Representation Theory of Measurement (RTM) and Classical Test Theory (CTT). I explain how both theories conceptualize measurement as a sharing of structure between an attribute under measure and the number series. To measure is to use numbers to represent intensity of the attribute, which is legitimate in virtue of shared structural properties. Measurability is thus understood as binary – to measure is to measure well; to measure poorly is to fail to measure at all. Validity is thus little more than an endorsement of the success of a measurement according to a measurement theory. I end this section by highlighting how the inadequacy of this simple picture of measurement in dealing with what Chang (2004) calls "the problem of nomic measurement" leads to an increased recognition of a need for more sophisticated *metaphysical* theorizing about measurement, which the construct validity program aimed to supply.

In section two, I discuss how validity theory struggles to balance theoretical concerns with practical limitations. The construct validity program sees validation as a process of using measurement to confirm a "strong theory" of some construct, which many areas in the social sciences do not have. In essence, the tension between "strong" and "weak" programs of construct validity reflects the same problem validity theory has always faced: we would like a valid test to both accurately reflect the structure of the world and be useful for some practical purpose, but attempts to meet both goals simultaneously have led to inconsistency. The construct validity program tried to prioritize the epistemic goal, but the pragmatic loss proved too high a price. Instead, advocates of the contemporary argument-based approach to validation choose to prioritize the pragmatic goal, thus marking a complete separation between measurement theory and validity theory.

In section three, I explore historical and sociological studies of measurement as a social activity. That is, there are important cases where the nature of a construct is fixed by pragmatic measurement choices, which are in turn guided by social, rather than epistemic, concerns. I argue that these circumstances call for measurement theories not built on the assumption that measurement happens only *after* the identification of a construct with an objective nature. Instead, we need to conceive of measurement as *creative* of the construct. The argument-based approach to validation provides important guidance by allowing validity theory to come before measurement theory.

In section four, I consider philosophical and methodological consequences of the observations made throughout this paper. Because it is possible for a test to be valid without measuring something that has test-independent ontology, the existence of valid tests alone should not count as evidence towards the test-independent ontology of whatever is under measure. I conclude with remarks on the relevance of my thesis to contemporary scientific debates around intelligence and predictive machine learning.

2.1 To Measure is to Measure Well

The field technically known as (Mathematical) Measurement Theory did not start as a theory of measurement, but as a theory of numbers. In 1887, the physicist Herman von Helmholtz published a paper titled *Zählen und Messen* (1887/1930), where he set out on a Kantian endeavor to found arithmetic in experience. He developed an axiom system of the positive integers based on the act of counting, and observed that discrete objects were not the only things "countable" in this sense. Attributes of objects, like length and weight, can also be counted, in the sense that they obey the axioms of counting. These attributes are deemed measurable and called *magnitudes*.

Although Helmholtz's theory of numbers did not gain much popularity[1], the idea that magnitudes and numbers obey the same set of axioms, which is what allows numbers to *repre-*

[1] Dedekind and Cantor opposed to Helmholtz's Kantian leaning, and Frege disliked his empiricism. See Darrigol (2003).

sent magnitudes, was taken up by mathematicians interested in structural similarity. With Hölder's axiomatization of *isomorphism* (1901/1996-7), an equivalence relation of structures, Helmholtz's theory of countability-as-isomorphism came to be seen as a theory of the measurability of attributes. According to the resulting Representational Theory of Measurement (RTM), an attribute is measurable just in case it is representable by numbers, and it is representable just in case it shares important structural properties with numbers. But which structural properties are important?

Norman Campbell, a physicist sometimes credited as the pioneer of modern measurement theory, argued that additivity was the defining feature of measurability, since additivity underlied the recursive definition of numbers (Campbell, 1938). This was a problem for psychologists: few, if any, psychological sensations admit a physical additive (or *concatenative*) procedure. In 1932, the British Association for the Advancement of Science (BAAS) appointed a committee to discuss the measurability of sensations. The committee remained undecided in its 1939-40 Final Report, with Cambell being its most outspoken and prominent voice against the measurability of sensations.

The most famous response to the Final Report was the psychologist S. S. Stevens' 1946 seminal treatment of scales. In a nutshell, Stevens argued that Campbell's insistence on additivity was dogmatic and unjustified. While additivity is an important property of numbers, there is no reason to prioritize it over other properties, such as orderliness. Attributes sharing other properties of numbers should also be deemed representable, so long as we clearly note the limitations of such representations.

Stevens' insight was quickly adopted and extended by mathematical psychologists to form the mature RTM. Despite its mathematical complexity, the basic ideas remained the same: an attribute is measurable just in case it shares certain properties with numbers, and to measure the attribute is to represent those properties with numbers. In other words, measurability is binary – to measure is to measure well, because to measure poorly is to fail to measure at all. Insofar as validity is about the success of measurement, validity theory was just a part of measurement theory.

In addition to RTM, the first half of the 20th century also bore witness to another foundational idea: mental testing. In 1904, Charles Spearman published a paper titled *"General Intelligence," Objectively Determined and Measured*, in which he reported a number of small scale studies on how children's judgments of pitch, brightness, and weight correlated with their school performance, and developed the method of factor analysis to "objectively measure" intelligence. He took great care in reviewing prior literature and reporting study procedures, but spent no time worrying whether what he had achieved was indeed a measurement of "general intelligence".

Despite this omission, Spearman's work on intelligence gave rise to both of psychometrics' major paradigms: Classical Test Theory and Item Response Theory. Although psychometrics has had little interaction with RTM[2], the underlying rationale can be put in similar terms: the dominant view at the time, which Spearman adopted and perpetuated, was that intelligence is an attribute which everyone possesses to some comparable degree. In other words, intelligence admits a total order. Since numbers also admit a total order, using numbers to represent intelligence is legitimate.

It is difficult to assess how much people challenged the assumption that intelligence is an objectively-existing, linearly-rankable attribute people possess. In 1923, E. G. Boring, experimental psychologist and Stevens' long-term collaborator, published an article defending the view that "measurable intelligence is simply what the tests of intelligence test", suggesting the presence of at least some skepticism. Similar to Spearman, Boring's main argument was

[2] The lack of interaction between RTM theorists and psychometricians has been called "the revolution that never happened" (Cliff, 1992), an observation concurred by others who have studied this period of history (Michell, 1999; Borsboom, 2005). My contention is that RTM, with its straightforward metaphysics and lack of a validity theory, no longer meets the need of contemporary testers. In fact, Item Response Theory, the dominant theory in modern psychometrics, is also largely ignored by field testers (Borsboom, 2006), for I believe to be the same reason.

that since people's relative rankings in score tend to remain stable across multiple mental tests, the aggregated ranking of test scores is objectively sound. However, it was not clear why the soundness of test score rankings should imply that intelligence is also rankable in a similar way and that the two rankings should correspond – both assumptions are necessary for the claim that test scores measure intelligence. It appears that Boring was at least somewhat conscious of this omission, since he provided something like a pragmatic argument for the fruitfulness of simply asserting that intelligence shares an underlying structure with test scores until further scientific observation tells us otherwise.

The problem Boring ran into was what Hasok Chang (2004) calls *the problem of nomic measurement*: measuring any unobservable quantity relies on knowing the precise mathematical relationship between the quantity in question and some observable indicator, which we can never know because the quantity under measure is unobservable. In this case, test scores measure intelligence only under the assumption that intelligence is totally orderable like test scores, which we can never test.

Like many others who ran into the same problem when struggling with measurement, psychometricians turned to operationalism for remedy. Classical Test Theory (CTT) was very operationalist in spirit. According to CTT, every person's test score is composed of two parts: the *true score*, representing the actual magnitude of the attribute under study, and a random *measurement error*. Each test, therefore, defines its own true score, much in the same way how each type of ruler defines its own length. Naturally, CTT runs into the same problems as operationalism[3] – how do we ever know if two tests test the same thing?

Concerns like these gave rise to a weak form of validity theory, *criterion oriented validity.* In its essence, a test is valid by this theory just in case we have reason to believe that we have crossed Chang's "nomic gap", i.e. that we have successfully established the relationship

[3] See Borsboom (2005) for a more detailed discussion on the parallel between operationalism and the limitations of CTT.

between the quantity under measure and the indicators. Of course, since it is actually impossible to cross the real nomic gap, criterion oriented validity was a rather weak concept. Its use was limited to cases when the quantity under measure was in fact observable, such as when we have “some other objective measure of that which the test is used to measure” (Bingham, 1937), or when we decide which quantity was plausibly under measure *after* the tests had already been done, such as when Guilford (1946) claims that “a test is valid for anything with which it correlates”.

Before moving on to discuss, in the next section, how this discontentment led to the construct validity program, I will briefly point out the role operationalism played in the development of RTM. Since Campbell saw concatenation procedures as the foundation for measurability, the topic of operation in measurement was around since the beginning of RTM. Stevens also embraced a substantive form of operationalism (Hardcastle, 1995), which was primarily motivated by his emphasis on the practical need for *agreement* among researchers. No matter how much disagreement there is about the unobservable nature of an attribute under study, everyone must agree on the observable, operationalized results. What scientists need to do is to figure out what, exactly, these agreed-upon results imply, which his theory of scales provides a framework for.

I believe Stevens’ operationalism provided the most consistent answer to the problem of nomic measurement: although we can never know what the real relationship is that holds between the (unobservable) quantity under measure and the (observable) indicators, we are free to make stipulations. As long as subsequent scientific theorizing respects the stipulative nature of this relationship, it is epistemically secure. It is my opinion that this is Stevens’ most important insight: “[s]cales are possible in the first place only because there is a certain isomorphism between *what we can do* with the aspects of objects and the properties of the numeral series” (Stevens, 1946, emphasis added). In other words, it is what we *do* with the measurement that gives it meaning. Unfortunately, this lesson is often overlooked by practitioners, much to the dismay of Stevens (1968) and subsequent measurement theorists (e.g., Suppés and Zinnes, 1963; see also Borsboom, 2005).

To briefly summarize: despite their peculiar lack of interaction, mathematical measurement theory (stemming from psychophysics) and psychometrics (stemming from intelligence research) share the same basic picture of measurement. According to this picture, an attribute under measure has some structural properties which it shares with the number system. To take measurement is to use numbers to represent the shared properties of the attribute. The problem of nomic measurement arises when researchers cannot be sure that the attribute does in fact possess these properties. Some form of operationalism was invoked to address this problem. Validity played little role in this narrative, partly because the challenge was not seen as one about measurement itself so much as it was about measurability, and measurability was understood as straightforwardly binary.

2.2 Validity: the Epistemic and the Pragmatic

Operationalism was not the only possible response to the problem of nomic measurement. Another is to invoke a form of robustness reasoning or, as Chang (2004) calls it, *overdetermination*. The idea is to conduct a number of studies on the same quantity under the same set of (unjustified) assumptions and using the corroboration amongst the results to justify those assumptions. It is clear that Spearman (1904) uses something like this: none of the children's test scores is, by itself, obviously an accurate measure of intelligence, but the fact that multiple such scores correlate with each other suggests that they all are.

However, as Chang has also argued, the method of overdetermination only works under *the principle of single value* – the belief that "a real physical property can have no more than one definite value in a given situation" (p.90, 2004). We can drop the word "physical" here, because the connection between this principle and the "realness" of the property is just as strong in discussions of mental properties.

More importantly, Chang (2001) points out that the principle of single value is not an empirically testable hypothesis. Instead, it is an *ontological principle* which a scientific community must accept for communication to be intelligible. It is what we use to make sense of empirical evidence, and therefore it cannot itself be supported or refuted by empirical evidence. If this principle can be justified at all, it must be justified theoretically.

The construct validity program, therefore, can be seen as an attempt at providing just this sort of justification. In response to the increased use of tests and the widespread confusion over how tests are supposed to work, the American Psychological Association, American Educational Research Association, and National Council on Measurements Used in Education jointly published a booklet titled *Technical Recommendations for Psychological Tests and Diagnostic Techniques* in 1954. It reviewed literature on a number of issues important to test users and provided guides to practice.

One such issue was validity. Around this time, the most commonly used conception of validity was criterion oriented validity as reviewed in section one. Another kind was *content validity*, which was simply looking at the wording of test items and deciding whether it sounds like it would measure the attribute it purports to measure. Both kinds are theoretically weak and practically arbitrary.

In the following year, two of the *Technical Recommendations*' authors published a treatise (Cronbach and Meehl, 1955) on the third kind of validity, *construct validity*, which was mentioned in the *Technical Recommendations* but not really discussed, apparently because some of the other authors were skeptical of this new idea. The paper provided a systematic treatment of what may be called a *metaphysics* of measurement, addressing questions such as what ontology must be assumed for measurement to be possible and what epistemology is supportable on this ontology.

The belief that measurement needed a metaphysics is itself significant. Recall the straightforward picture of section one: an attribute has some properties which we can determine through scientific studies; the number system has some properties which we understand from mathematics. Measurement is possible just in case there is appropriate overlap between the two sets of properties. Metaphysics was hardly relevant.

But metaphysics was everywhere relevant once we realized that empirical evidence alone cannot support key assumptions like the principle of single value: how do we know that, when two tests do not correlate, it is because one test has low validity rather than that they measure different things? Or, conversely, why do we think tests that highly correlate with each other measure one attribute rather than a number of closely connected attributes? If these questions have answers at all, they would not come to us through simple empirical observation. Robust theorization is required.

Robust theorization is what the construct validity program aims to provide. According to this picture, the nature of an unobservable quantity under measure comes in the form of a robust theory. Following logical empiricism, this theory takes the form of a "nomological net" (Cronbach and Meehl, 1955) which specifies how this quantity causally relates to other established quantities and, ultimately, traces to empirical observations. In typical logical empiricist fashion, the quantity is *real* in the sense that its legitimacy must ultimately be traceable to experience, but it is also *instrumental* in the sense that it is not discovered in the world, but *constructed* as a way to make sense of the world.

Also in typical logical empiricist fashion, the theory of constructs appears to have presented a brilliant solution to seemingly-unresolvable problems. Assumptions like the principle of single value and the (unobservable) isomorphism between the construct and its indicators are not swept under the rug like they were in early psychometrics. Instead, they are provided a natural space in a comprehensive ontology, with the promise that they can be supported like any other scientific claim: through a combination of empirical evidence, inference to the best explanation, and inferential holism.

To be clear, the belief that measurement should be about "real" attributes that share properties with the number series did not change with the advent of construct validity. In Loevinger's words, "[t]he basic concept is that of the construct validity of the test, the degree to which it measures some trait which really exists in some sense" (1957). Instead, the real insight of the new theory is the recognition that the validity of a test – that the test has done what it is created to do – is not directly observable, but must be inferred through circumstantial evidence.

Central to this insight is a more principled view of the relationship between a test's *epistemic content* and its use. By epistemic content, I mean what the test supposedly tells us about the attribute being measured. In the language of RTM and early psychometrics, a (successful) test of an attibute uses numbers to represent facts about that attribute. For example, calling one length "2 meters" and another "1 meter" is to express the fact that concatenating two rods of the second length would yield a length equivalent to the first. The test says something about the world that may be true or false. On the other hand, test results can also be used for some practical purpose. For example, knowing the distance between two cities can help me judge how much fuel I need in my car.

Prior to the construct validity program, researchers differed radically in how they understood the relationship between the epistemic content of a test and its use. Guilford (1946) distinguished "factorial" and "practical" validities, where factorial validity was the extent to which a test tracks the underlying factor or construct, and practical validity was how useful a test was to testers. We have seen how Boring argued that we could operationally define intelligence as "what the tests test", thus citing the usefulness of these tests as evidence for the legitimacy of their empirical content. Flipping Boring's logic on its head, Anastasi (1950) argued that, because we would not want to operationally define intelligence in this way, validity should always be about the usefulness of tests and never their empirical content, lest it "mislead us into the belief that anything external to the tested behavior has been identified" (p.75).

After the construct validity program, the idea that the usefulness of a test can serve as evidence for its empirical content without invoking a naive form of operationalism was widely accepted. Since a construct is defined through a complex "nomological net" that ultimately, but not immediately, depends on experience, the pragmatic success of its tests naturally constitute a majority, but not the entirety, of the evidential support for the epistemic success of the theory. Validity is about the truth (or empirical adequacy) of a theory of a construct. Measurement theory is important insofar as it tells us how we should interpret measurement results for the sake of validation, but it no longer subsumes validity theory.

The measurement theory that best fits the construct validity program is Item Response Theory (IRT; also called latent variable theory). Like CTT, IRT is traceable to Spearman's 1904 papers on intelligence, but is less operationalist and more mathematically sophisticated than CTT. The basic idea of IRT is that a small number of latent variables are causally responsible for a person's performance on a large number of tests. An IRT model can be seen as a specific articulation of the broader theory under study, with validity being a measure of how well it is supported by evidence.

It is perhaps unfair to say that the construct validity program "failed", but rather that it fell short on almost all of its promises and was abandoned. Testers soon realized that the kind of robust "nomological net" of constructs that lay at the heart of construct validation is almost never feasible in practice. When law schools test their recruits for "academic promise" or companies test their customers for "service satisfaction", they are unlikely to have anything close to a well-founded theory of a construct in mind.

It does not look as if they need one, either. Testers can agree that one test is better than another at predicting law school GPA or future sales without agreeing on why. It may be that one test most accurately captures the construct of "customer satisfaction", or some other construct that is a common cause for both test scores and sales, or any other underlying causal structure for which there is no evidence. What they do have evidence for is how well a test does its pragmatic job. And, for most practical purposes, this is enough.

The fact that strong theories of constructs are often neither easy to acquire nor practically necessary led to a division between the strong and the weak programs of construct validity (Cronbach, 1988; Kane, 2001). The strong program is what Cronbach and Meehl (1955) intended – validity theory is the confirmation of a robust "nomological net" of a construct based on extensive evidence and argumentation. When no such theory is available, researchers meet the demand for test validation by gathering "an unordered array of correlations with miscellaneous other tests and demographic variables. Some of these facts bear on construct validity, but a coordinated argument is missing" (Cronbach, 1980). This is the weak program.

In simple terms, the story can be summarized as follows. Our original goal is to assess an attribute (e.g., customer satisfaction) in order to perform a task (e.g., predicting sales). A test can help with this task by accurately measuring the attribute. Assessing the accuracy of a test in tracking the attribute has been the main motivation behind validity theory up until this point. However, the strong versus weak program debate highlights the fact that it is much more difficult to judge whether a test has succeeded in measuring the attribute than whether it has succeeded in helping us perform the task. Since our original goal is to perform the task rather than to assess the attribute, chasing success in measurement seems to be an unnecessary intermediary. By 1988, even Cronbach has conceded: "Still, the weak program has some merit".

This realization motivated the reconceptualization of validity not as a judgment on the quality of the relationship between a test and an attribute, but as the extent to which a test can help with a particular task. In an often-cited chapter of the third edition of *Educational Measurement*, Messick (1989, p.13, emphases original) opens with the following definition:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment.

Although specific aspects of Messick's subsequent analysis of validity were disputed by other theorists, the basic idea that validity is about the "adequacy and appropriateness" of "inferences and actions" remained. The 2014 edition of *The Standards*, for example, declares that, in validation, "[i]t is the interpretations of test scores for proposed uses that are evaluated, not the test itself" (p.11). In other words, whether a test is valid no longer depends on whether it stands in the right kind of relationship with an attribute in the world, but on what testers plan to do with the test results. Validity theory is no longer about measurement.

The view that validity is an evaluation not of how tests measure but of how tests are used underlies the argument-based approach to validation (Shaw and Crisp, 2011; Kane, 2013a). As I have shown, this view arises not because of its theoretical superiority, but from practical necessity. Indeed, the attitude that, when theory and practice come into conflict, it is theory that should yield seems widespread among testers. For example, Shepard (1997) considers a case where pre-med students prioritize science classes over humanities as a way to increase MCAT scores, thus making the MCAT no longer an adequate test for identifying students who are more likely to succeed in medical school. The validity of a test appears to have changed over time. Shepard goes on to argue that this example should not be taken as a case against pragmatically defined validity, but as a call for reexamination of the relevance of the construct tested by the MCAT. In other words, when the test-assisted inference comes into conflict with the theory of construct, it is the latter that we need to modify.

More recently, Borsboom and collaborators have raised another theoretical challenge, namely that validity should be about the truth of ontological claims: “The attribute to which the psychologist refers must exist in reality; otherwise, the test cannot possibly be valid for measuring that attribute” (Borsboom et al., 2004, p.1063). To use Loevinger’s words again, a test is valid just in case it “measures some trait which really exists in some sense”. Part of their reason is that Item Response Theory, the dominant theory of measurement in psychometrics, calls for a realist ontology in which an actually existing attribute is causally responsible for observed test scores (Borsboom et al., 2003). In other words, if validity is going to be an evaluation of measurement theory at all, it must take a stance on the reality of the target of measurement.

Philosophers have raised similar concerns. For example, Alexandrova and Haybron (2016) point out that, while the very idea of construct validity relies on a theory of construct, contemporary practices of validation tend to be theory avoidant. In Cronbach’s (1980) words, “a coordinated argument is missing.” On the other hand, Stone (2019) advocates for a distinction between construct legitimacy and construct validity, where legitimacy concerns the realist ontology and validity concerns the success of measurement.

However, while these challenges all have important theoretical upshots, the motivation for the argument-based approach has never been theoretical. “In many testing situations (including most high-stakes contexts), talk of Truth seems hollow”, Kane (2013b) explains. “I am more pragmatic (with a small “p”). I am concerned about what can reasonably be claimed on the basis of test scores in the current context.”

Neither comprehensive nor theoretically elegant, Kane's response nevertheless lies at the heart of the issue. It is not that strong theories or realist ontologies are undesirable; they are simply too much trouble for their worth.

To summarize, in this section, I reviewed how validity theory struggled to balance two goals of testing: the epistemic goal of accurately describing a construct that exists in the world and the pragmatic goal of using the test to accomplish some practical task. The construct validity program treats the epistemic goal as primary. However, despite its many theoretical advantages, prioritizing the epistemic goal proved both unfeasible and unnecessary. In response, the contemporary argument-based approach to validation treats validity as the evaluation of an inference or action that is made on the basis of a test. Since the existence of a construct is no longer central to the new paradigm of validation, validity theory and measurement theory become independent fields of study.

2.3 Theory Avoidance: Bug or Feature?

In section one, I reviewed how measurement theory started as a theory about the relationship between a test (the instrument of measurement) and an attribute or construct[4] (the target of measurement). In section two, I reviewed how validity theories needed to balance the epistemic goal of making sure a test accurately captures the construct and the pragmatic goal of allowing test users to perform a task. The construct validity program attempted to achieve such a balance by adopting a logical empiricist framework in which a construct is defined by its causal connections. Unfortunately, its heavy reliance on strong theories makes construct validation unfeasible in many practical contexts. In its place, the argument-based approach to validation holds that a test is valid *in a context* just in case its results can be

[4] As explained in section two, the term "construct" was created primarily to emphasize the role of a "nomological net" in reference to a logical empiricist ontology. This particular association of the term is dropped in contemporary discussions of validity. Therefore, I will use "construct" and "attribute" interchangeably to mean the aspect of the world supposedly measured by a test.

used to support a specific inference or action. The argument-based approach is sometimes criticized for its disconnection with measurement theory. In this section, I argue that the new validity theory's detachment from the theory of constructs, far from being a weakness, makes it better situated at accommodating new understandings of measurement.

In his analysis of the history of quantification, Porter (1996) argues that, while people tend to see quantification as faithful documentation of objective facts, the real story is the exact opposite: quantification is invoked to cope with a reality that is neither objective nor 'documentable'. Many forms of social statistics, Porter explains, were created in response to social pressures of expanding societies across greater geographical distances. For example, to identify discrepancies in standards of living across regions for the purpose of focused development, one must be able to measure "standard of living" of regions in a way that makes them comparable.

Because numbers underly the most obvious way of cross-comparison, phenomena more easily describable with numbers were prefered. The weight of a heap of grain was valued over its quality, which had to be judged 'subjectively' by a skilled inspector. Insurance companies used to personally interview potential clients to judge how much they should charge for a life insurance. Now they use numbers such as age, number of hospital visits, and annual income to make the same judgement, allowing information to be more easily summarized and compared.

In other words, what information gets included in a measurement is a decision dictated by practical, rather than epistemic, concerns. The square footage of a house tracks the quality of the house, which tracks the quality of life of a family. But this does not mean that home size is an epistemically privileged indicator for standard of living. That is, the decision to measure standard of living by home size is not motivated by the belief that home size shares more structural features with standard of living than other indicators and therefore is better situated at revealing the true nature of this construct.

This is not to say, of course, that the indicator of choice does not in fact share sufficient structural features with the construct to count as successful measurement in the representational sense. But this view becomes very difficult to hold once we consider the fact that questions about the structural features of the construct never come into play in decisions of whether or how to measure it. When economists use numbers to analyze and “measure” wealth inequality, the possibility that wealth inequality might not admit a total order is never at the center of debate.

Central to historians’ analysis of measurement practice is the observation that the nature of a construct often follows, rather than dictates, the decision over how to measure it. For example, one major victory of 19th century feminist movement for property rights was to have housework count as work and, in doing so, reconceptualized women’s role in the labour market (Siegel, 1994). Espeland and Stevens (1998) have argued that capitalist standardization, or the idea that everything can be priced along the same axis, is at least partially responsible for extreme global wealth inequality. Many of forestry’s basic conceptual frameworks were developed based on the need to classify forests into ‘useful’ and ‘not useful’ parts for the purpose of calculating lumber yields (Lowood, 1990; Scott, 1996).

These observations make it difficult to accept, as measurement theorists have, that measurability is dictated by the objective nature of the construct – that some constructs are measureable in virtue of possessing certain properties while other constructs are not – and that is all measurability is about. Instead, everything is measureable as long as there is a practical need to measure it. The question then becomes: what are we doing when we measure something, if not using numbers to describe its objective nature?

A radical answer, provided by Comaroff and Comaroff (2006), is that we are *inventing facts*. In their paper, Comaroff and Comaroff document the political and social dynamics surrounding the development of crime statistics in South Africa. On the one hand, crimes happen in the world, and so there are facts of the matter whether certain measures of crimes are accurate or inaccurate. In this sense, crime statistics are *factual.* On the other hand, decisions that crucially define what crime statistics will end up looking like – decisions concerning how to classify crimes, who to ask, how to ask, how to count multiple aggressions in the same scenario, etc. – are made not on the basis of how a theory of crime understands the construct of crime rate, but for political, social, and practical reasons. In this sense, crime statistics are *invented.*

This dynamic can be difficult to discern. No matter how little epistemic consideration underlied its foundation, a system of measurement tends to develop a facade of objectivity which makes future users unable to challenge its authority. In a sociological study of the creation of human rights indicators for the United Nations, Merry (2016) argues that, although indicators are often presented as democratically revisable, the fact that the creation of indicators changes the very language by which we use to talk about the world means that, in practice, they are immune to outside scrutiny. Consequently, contentious decisions made by experts who were present at the start can have consequences long exceeding the amount of theorization undertaken at the time. In Merry's words (2016, p.21):

> For example, in order to measure violence against women, throwing acid in the face of one's wife in Bangladesh must be equilibrated to shooting a domestic partner in the United States. This intellectual, interpretive work is shaped by the politics of expertise and participation that determine how quantitative knowledge is developed and by whom.

To use a familiar philosophical metaphor, nature starts without joints. Joints are carved into nature through pragmatically motivated measurement efforts, which then gives off the appearance that measurement has always "carved nature at its joints".

This idea of measurement practice defining a construct shares important features with Hasok Chang's theory of epistemic iteration. In his analysis of the history of thermometry, Chang

(2004) explains how the concept of temperature was shaped by practically motivated measurement choices in an iterative process: first, a pre-theoretic understanding of coldness and warmth serves as a starting point for building thermoscopes. Next, development in theories of thermometry changes how we understand temperature. On the one hand, the measurement target has shifted over time and so ealier measurement results should no longer be seen as valid. On the other hand, many of the key insights in new conceptualizations of the target stem from observations of past, now-outdated measurement attempts. The paradoxical observation is that, while new theories of temperature made old theories of measuring it invalid, the former is also founded upon the latter.

There are important disanalogies, too. Part of Merry's thesis is precisely that measurement developments often lack the iterative feature Chang describes, leading to an inflexible and, therefore, undemocratic process. Chang's analysis of temperature also highlights the importance of a pre-theoretic but widely agreed upon conception of temperature, a condition disputed by Comaroff and Comaroff in the case of crime statistics.

What this means is that there are important nuances to be studied in these measurement instances. Since they share the feature where the practice measurement shapes the theory of the construct, any measurement theory starting with the assumption that measurement is simply a correspondence relationship between a construct and an indicator will be inadequate as a framework of analysis.

This is why the argument-based approach to validation, though not a theory of measurement, provides a better starting point for measurement cases described in this section. In particular, a test can be valid in the sense that users are as justified in using its results to make inferences as they are any other valid tests, without having independent reasons to believe that this test has succeeded in capturing the objective nature of a construct that existed before anyone tried to measure it. According to this theory of validity, on the one hand, validity is pragmatic, because the kind of reasons one would give to judge validity are based on what a test can help us do. On the other hand, validity is not 'merely' pragmatic, because testers did not *choose* to ignore the epistemic content of a test that is there. The epistemic content of a test is not 'there' in any robust sense. Validity theorists are not choosing to ignore it if they cannot choose to not ignore it, because it does not exist yet. Instead, the epistemic content of a test is created through a process that starts *after* some tests are pronounced valid.

It remains true that the argument-based approach to validation is a theory of validity, not a theory of measurement, and therefore does not provide us with a story about how an indicator is supposed to relate to the construct. Just as how Item Response Theory (a measurement theory) fits well with the construct validity program (a validity theory), we can develop new measurement theories that fit well with the argument-based approach to validation. It is beyond the scope of the present paper to develop such a theory, except to note that the first step in doing so is to give up on the doctrine that valid measurement must be about attributes that exist independently of how they are measured.

2.4 Measuring the Non-Existent

So far I have argued that valid measurement does not have to be about attributes that objectively exist before they are measured. It is worth noting that my reasons, like those of many validity theorists, are primarily pragmatic. It is not that there is anything philosophically undesirable with the classical, clean picture, where we pick out an attribute in the world, axiomatize its structure, and see whether an isomorphism exists between it and the number series. This picture is simply unrealistic in many (but perhaps not all) cases where we need to take measurement. When it comes to deciding what to measure, 'measurability' gives way to pragmatic concerns.

It is also worth noting that, although theories of validity differ in how they understand the *meaning* of validity, the actual *evidence* presented to support a validity claim is largely the same. A test on X is valid if it correlates with either other tests on X or tests on constructs thought to be causally related to X. The SAT is valid if it correlates with college GPA. According to criterion oriented validity, this is because college GPA is an accurate measure of what the SAT tries to measure. According to the construct validity program, this is because predicting college GPA is an important part of a theory of a construct for which the SAT measures. According to the argument-based approach to validation, this is because the SAT is designed to help with college admissions, and correlating with college GPA suggests that the SAT is up for the task. That the correlation is evidence for validity is always true; the interpretation of why that is changes.

What this means is that we should be careful with what we conclude from the pronouncement of validity. Tests are declared valid, used broadly, and incorporated into our understanding of the world based on a variety of measurement and validity theories. Since these theories disagree over what the necessary preconditions are for a test to be valid, a pronouncement of validity should not be seen as evidence for these preconditions. Put in another way, inference to the best explanation is not a good inference unless there is widespread agreement over what the best explanation is.

The thesis that we can legitimately consider a test to be valid without committing to the existence of a test-independent construct has important consequences in contemporary debates around the ethics and epistemology of measurement-driven science. While it is beyond the scope of this paper to provide full analyses of such cases, I will briefly sketch the relevance of my thesis in two examples: intelligence research and predictive machine learning.

As reviewed in section one, a large part of the field now called psychometrics was born out of intelligence research and has a troubling history with racism and colonialism. Nevertheless, it may be argued that the construct of intelligence is itself value neutral and that, once we purge the troubling part of the theory, the rest will prove scientifically useful. Based on my argument in this paper, there are at least two complications with this view. First, since *what* intelligence is depends on *how* it is measured, which in turn depends on *why* it is measured, purging the pragmatic context around intelligence may simply destroy the construct altogether. Second, the correlation-based validity claims typically made in intelligence testing is insufficient at supporting many ontological claims associated with intelligence as a construct, and so there may not be a construct to save to begin with.

As in the case of crime statistics discussed in section three, many of the behavioural observations underlying judgments of intelligence are real. Some people learn mathematics slower than others. In this sense, many intelligence-related claims correspond to facts. However, the idea that intelligence is a construct goes beyond these claims. It includes claims such as 'reading speed and mental arithmetic are two behavioural manifestations of the same physiological cause'. These claims depend crucially on (1) what kinds of tests are used, and (2) the assumption that the validity of these tests is evidence for the test-independent (e.g. physiological) existence of a single construct. As I have argued in this paper, (1) lacks epistemic foundation and (2) is unjustified.

Moving to the second case, although predictive machine learning is not typically seen as a measurement-driven field, I believe a similar dynamic, and hence similar dangers, also apply. Predictive machine learning refers to the practice of using data to train algorithms to help with future decision making. Much research has been done on the danger of algorithmic bias, which is when algorithms lead to biased, and often discriminatory, decisions. Algorithmic bias can occur for a number of reasons: bias in the training data set, bias in the algorithm itself, sample misrepresentation, etc. However, researchers are increasingly recognizing that algorithmic bias may not be preventable: it is often not a sign that something has 'gone wrong', but a reflection of the intrinsic weakness of a style of reasoning. For example, Johnson (2020) argues that, since machines ‘learn’ by associating proxy attributes to target classifications, stereotype reasoning is an inherent part of learning.

The success or failure of machine learning is often couched in terms of success or failure in picking out ‘real patterns’ in the population from data. The idea is that successful predictions occur when the relationship between predictor variables and outcome variables is causal or material, rather than merely statistical or correlational. In a sense, data are measures of the world. The fact that data are useful in making predictions is a testament of validity. The claim that successful predictions by data-trained algorithms must mean that the data have captured ‘real patterns’ is a claim about validity implying ontology. I have argued that this claim is unjustified. Consequently, we must rethink the connection between predictive success and ontological reality.

In conclusion, I have argued that there are many reasons to accept the framework where validity theories can be independent from, and prior to, measurement theories: measurement theories are unable to solve philosophical problems about measurement anyway; measurement-based validity theories are pragmatically too limited; and there are important real world measurement cases that cannot be adequately addressed by existing, measurement-based validity approaches. Instead of seeing this as a shortcoming of present-day validity theory, however, I have argued that the measurement-independent aspect of the argument-based approach to validation opens up a new and much-needed space for testers to reexamine the nature of valid measurement and the relationship between measurement and ontology.

Chapter 3

A Statistical Learning Approach to a Problem of Induction

Hume's problem of induction can be analyzed in a number of different ways. At the strongest, it denies the existence of any well justified assumptionless inference rule that is ampliative. At the weakest, it challenges our ability to consistently apply, in practice, any such rule that might exist. This paper examines an answer to the latter problem in the context of statistical learning theory and argues for its inadequacy.

The particular problem of induction discussed in this paper concerns what Norton (2014) calls a formal theory of induction, where "valid inductive inferences are distinguished by their conformity to universal templates" (p.673). In particular, I focus on the template that is often called *enumerative induction*. An inductive argument of this type takes observations made from a small and finite sample of cases to be indicative of features in a large and potentially infinite population. The two hundred observed swans are white, so all swans are white. Hume argues that the only reason we think a rule like this works is because we have observed it to work in the past, resulting in a circular justification.

Nevertheless, this kind of inductive reasoning is vital to the advancement of a scientific understanding of nature. Most, if not all, of our knowledge about the world is acquired through the examination of only a limited part of the world. The scientific enterprise relies on the assumption that at least some of such inductive processes generate knowledge. With this assumption in place, a weak problem of induction asks whether we can reliably and justifiably differentiate the processes that do generate knowledge from the ones that do not. This paper discusses this weak problem of induction in the context of statistical learning theory.

Statistical learning theory is a form of supervised machine learning that has not received as much philosophical attention as it deserves. In a pioneering treatment of it, Harman and Kulkarni (2012) argue that one of the central results in statistical learning theory – the result on Vapnik-Chervonenkis (VC) dimensions – can be seen as providing a new kind of answer to a problem of induction by providing a principled way of deciding if a certain procedure of enumerative induction is reliable. The current paper aims to investigate the plausibility of their view further by connecting results about VC dimension in statistical learning with results about NIP models in the branch of logic called model theory. In particular, I argue that even if Harman and Kulkarni succeed in answering the problem of induction with the VC theorem, the problem of induction only resurfaces at a deeper level.

The paper is organized as follows: section 1 explains the relevant part of statistical learning theory, the VC theorem, and the philosophical lessons it bears. Section 2 introduces the formal connection between this theorem and model theory and proves the central theorem of this paper. Section 3 concludes with philosophical reflections about the results.

3.1 Statistical learning theory

The kind of problems that is relevant for our discussion of VC dimensions is often referred to as classification problems that are irreducibly stochastic. In a classification problem, each individual is designated by its k-many features such that it occupies somewhere along a k-dimensional feature space, χ. The goal is to use this information to classify potentially infinitely many such individuals into finitely many classes. To give an example, consider making diagnoses of people according to their test results from the k tests they have taken. The algorithm we are looking for needs to condense the k-dimensional information matrix into a single diagnosis: sick or not. The algorithm can be seen as a function $f : \chi \to \{0, 1\}$, where 1 means sick and 0 means not. For reasons of simplicity, I will follow the common practice and only consider cases of binary classification[1].

By "irreducibly stochastic", I mean that the target function f cannot be solved analytically. This might be because the underlying process is itself stochastic – it is possible for two people with exact same measures on all tests to nevertheless differ in health condition – or because the measurements we take have ineliminable random errors. This means that even the best possible f will make some error, and so the fact that a hypothesis makes errors in its predictions does not in itself count against that hypothesis. Instead, a more reasonable goal to strive towards is to have a known, preferably tight, bound on the error rate of our chosen hypothesis.

What makes this form of statistical learning "supervised learning" is the fact that the error bound of a hypothesis is estimated using data points whose true classes are known. Throughout this paper, I will use D to denote such a dataset. D can have any cardinality, but the interesting cases are all such that D is of finite size. Recall that the feature (or attribute)

[1] Results for binary classification problems generalize straightforwardly to finite classification problems. They also generalize, with some caveats, to approximation problems with less-than-perfect precision, such as when two numbers that agree up to the fifth decimal place are considered practically identical.

space χ denotes the space of all possible individuals that D could have sampled, so that $D \subset \chi$. I understand a hypothesis to be a function $h : \chi \to \{0, 1\}$. A set of hypotheses $\mathcal{H}$ is a set composed of individual hypotheses. Usually, the hypotheses are grouped together because they share some common features, such as all being linear functions with real numbers as parameters. This observation will become more relevant later.

One obvious way of choosing a good hypothesis from $\mathcal{H}$ is to choose the one that performs the best on D. I will follow Harman and Kulkarni (2012) and call this method enumerative induction, for it bears some key similarities with Hume's description of the observation of swans. This method is inductive because it has the ampliative feature of assuming that the chosen hypothesis will keep performing well on individuals outside of D. The question we are interested in is: how do we know if this generalization is true? What justifies the claim that the hypothesis performs well on D will perform well outside of D too? The answer that will be examined in this section and throughout the rest of the paper is that we know this claim to be true when we are in a situation where $\mathcal{H}$ has finite VC dimension, and the VC-theorem justifies this claim.

To define the error rate of a hypothesis, recall the "ideal function" f mentioned in the introduction. Recall also that f classifies individuals from χ into $\{0, 1\}$, and f is imperfect. Nevertheless, since the process from χ to the classes is irreducibly stochastic, f is as good as we can hope for. Therefore, f will serve as our standard for the purpose of calculating the error rate of a hypothesis. Note that the hypotheses we are assessing are all from $\mathcal{H}$, our hypothesis set, but f need not be in $\mathcal{H}$.

Suppose D is of size N, and $x_1, \ldots, x_N \in D$. For each $h \in \mathcal{H}$ and $i \in [1, N]$, consider the random variable $X_i : \chi^N \to \{0, 1\}$ defined by

$$X_i\big(h(x_1, \ldots, x_N)\big) = \begin{cases} 1 & \text{if } h(x_i) \neq f(x_i), \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

Intuitively, $X_i = 1$ if the hypothesis we are evaluating, h, gives a different (and hence wrong) verdict on x_i than the target function f, and 0 otherwise. Assume $X_1, \ldots, X_N$ are independent, which is to say that making a mistake on one data point does not make it more or less likely for h to make a mistake on another one. This is typical if D is obtained through random sampling. Further assume $X_1, \ldots, X_N$ are identically distributed, which means that the expectations of all of these variables are identical, or $EX_i = EX_j$ for all X_i and X_j. This allows the error "rate" of h across multiple data points to be meaningfully computed.

Let $\overline{X} = \frac{1}{N}(\sum_{i=1}^{N} X_i)$, which is the measured mean error, and $\mu = E\overline{X}$, which is the expected mean error. I will follow Abu-Mostafa et al. (2012) in calling the former the *in-data error*, or E_{in}, and the latter *out-data error*, or E_{out}. To flesh out the relationship between these two values more clearly, we define

$$E_{in}(h) = \overline{X} = \frac{1}{N}\sum_{i=1}^{N}[\![h(\mathbf{x}_i) \neq f(\mathbf{x}_i)]\!] \tag{3.2}$$

$$E_{out}(h) = \mu = \mathbb{P}_N(h(\mathbf{x}) \neq f(\mathbf{x})) \tag{3.3}$$

Intuitively, the in-data error is the evidence we have about the performance of h, and the out-data error is the expectation that h will hold up to its performance. The amplification comes in when we claim that E_{out} is not very different from E_{in}. I will call the difference between E_{in} and E_{out} the *generalization error.*

For any single hypothesis, and for any error tolerance $\epsilon > 0$, Hoeffding (1963, p.16) proved a result called the *Hoeffding inequality*[2], which states that, under the assumption that the error rate for each data point is independent and identically distributed, we have (in the notations introduced above)

$$\mathbb{P}_N(|E_{in}(h) - E_{out}(h)| \geq \epsilon) \leq 2e^{-2\epsilon^2 N} \tag{3.4}$$

This inequation says that the probability of having a large generalization error in the assessment of a single hypothesis is bounded by $2e^{-2N\epsilon^2}$, which is a function of the size of the dataset, N, and the error tolerance ϵ.

Once we establish a bound in the case of a single hypothesis, we can get a similar bound for finitely many such hypotheses. The reason we cannot simply apply the Hoeffding inequality to our preferred hypothesis is that the bound is generated with respect to a *random* dataset. That is, we need to decide on which hypothesis to evaluate *before* we generate the data. If this condition is violated, then the Hoeffding inequality is no longer valid. In the context of supervised statistical learning, the dataset is needed for us to decide which hypothesis we "prefer". Consequently, we cannot use the same dataset to bound the generalization error of this preferred hypothesis. To get around this problem, we need to make sure that *all* hypotheses in the set have low enough generalization errors, so any one hypothesis we pick out will, too.

Since we assume that the error rate of one hypothesis is independent of another, the probability of any of the finitely many hypotheses we are considering having a large generalization error is just going to be the union of the probability of each one of them does. In symbolic

[2] see also Lin and Bai 2010, p. 70, and Pons 2013, p. 205

form, suppose there are $1 \leq M < \infty$ many hypotheses in $\mathcal{H}$, then we have (Abu-Mostafa et al., 2012)

$$\mathbb{P}(\max_{h \in \mathcal{H}} |E_{in}(h) - E_{out}(h)| \geq \epsilon) = \mathbb{P}(\exists h \in \mathcal{H} | E_{in}(h) - E_{out}(h)| \geq \epsilon) \leq 2Me^{-2\epsilon^2 N} \tag{3.5}$$

While this bound may seem "loose", it serves our purpose when we have a reasonably small M or a reasonably large N.

This simple calculation becomes tricky, however, when $\mathcal{H}$ contains infinitely many hypotheses. If we replace M with infinity, then the upper bound stops being a bound, because $2Me^{-2\epsilon^2 N}$ grows to infinity as M does. This is where the VC dimension of $\mathcal{H}$ comes to play.

To understand the role of VC dimensions, consider a "verdict tuple" $(h(\mathbf{x}_1), \cdots, h(\mathbf{x}_N))$ for a hypothesis h on a sample D of size N. For a binary problem where the available classes are 0 and 1, the verdict tuple will have N entries of 0s and 1s, one for each of the N elements in D. If some hypotheses agree with each other on the classification of every data point, then their verdict tuples would be identical. Further define

$$\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N) = \{(h(\mathbf{x}_1), \cdots, h(\mathbf{x}_N)) \mid h \in \mathcal{H}\} \tag{3.6}$$

which is the set of all verdicts given by $\mathcal{H}$ on dataset D. Since two or more hypotheses may agree on all verdicts, the cardinality of the set of verdicts may be much smaller than the cadinality of $\mathcal{H}$. Furthermore, how many different verdict tuples $\mathcal{H}$ generates may also change with different datasets, since hypotheses that agree on one set of N elements may not agree on another set of N elements. Define

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \cdots, \mathbf{x}_N \in \chi} |\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)| \tag{3.7}$$

as the max number of different verdicts $\mathcal{H}$ can generate from any dataset of cardinality N.

If all possible classifications of D have been represented in $\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)$, then we have $m_{\mathcal{H}}(N) = 2^N$. When this happens, we say that the hypothesis set $\mathcal{H}$ *shatters* the dataset D. Define the *VC dimension of* $\mathcal{H}$ to be the maximum N such that $m_{\mathcal{H}}(N) = 2^N$. In other words, it is the maximum number N such that there exists a dataset D of size N that is shattered by $\mathcal{H}$. If $m_{\mathcal{H}}(N) = 2^N$ holds for all N, then we say the VC dimension is infinite. Let's call a hypothesis set $\mathcal{H}$ VC-learnable if it has finite VC dimension.

Very roughly, the VC dimension of a hypothesis set tracks the maximum number of hypotheses that are still distinguishable from each other with respect to their verdicts on data. This means that, if we consider any more hypotheses, some of them will always agree with some others on all of the classifications they give to all possible data points, and so if one has low generalization error, the others will, too. The VC generalization bound is given as follows (Abu-Mostafa et al., 2012, p.53)

$$\mathbb{P}_N[\![\big(E_{out}(h) - E_{in}(h)\big) \leq \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}]\!] \geq 1 - \delta \tag{3.8}$$

where δ is the uncertainty tolerance. If $\mathcal{H}$ has an infinite VC dimension, then no such upper bound can be found. Notice that, holding everything else equal, increasing N brings the right-hand side down, which means that increasing data size allows us to make a better estimate of E_{out} with the same uncertainty tolerance. This means that, when $\mathcal{H}$ is either finite or has finite VC dimension, we can justifiably claim enumerative induction to be a reliable process that can pick out a good hypothesis from $\mathcal{H}$.

What makes this theorem especially powerful is not just that it shows how the error rates converge in the limit, but also that the convergence is uniform. What is practically useful for statisticians is not so much that, if we have infinite data, we can figure out the true error rate of our hypothesis, but that, as soon as we know how many data points we have and the

VC dimension of $\mathcal{H}$, we know precisely how confident we should be of our estimation of the error rate.

In what sense does this theorem answer a problem of induction? According to the analysis in Harman and Kulkarni (2012), this theorem defines precise conditions (i.e., ones where $\mathcal{H}$ has finite VC dimension) under which a particular inductive method (i.e., supervised learning in classification problems) is reliable. To the extent that we are concerned with the "easy" problem – the practical problem – of induction, the VC theorem does seem to provide a kind of answer we are looking for. In the next section, I challenge the applicability of this answer. In particular, I show that we can never know in general if we are in a situation where the above answer is applicable.

3.2 Finiteness of VC dimensions is uncomputable

A preliminary observation about the finite-VC requirement is that we do not have a good grasp of what it tells us about these hypotheses. What is the intuitive or scientific difference between these two sets of hypotheses such that one has finite VC dimension and the other does not? There is no straightforward answer to this question. It seems to be a brute fact that some hypothesis sets behave nicely and others do not. To put this point more concretely, we know that polynomial functions with arbitrarily high degrees have finite VC dimension, whereas the set of formulas with the sine function has infinite VC dimension. What is the difference between them? If we have a problem that can be reasonably formulated as polynomials or with a sine function, do we have good principled reasons why we should formulate it in one way rather than another?

Surprisingly, model theory in logic might help shed light on this question. It turns out that the concept of *NIP* – the *Not-Independent Property* – theories corresponds to the class of hypothesis sets with finite VC dimensions. A theorem provably equivalent to the VC theorem was independently proven by the model theorist Shelah about these NIP theories and the corresponding NIP models in a way that makes results about VC-learnable sets and results about NIP models intertranslatable. This connection was first recognized by Laskowski (1992).

In the previous section we discussed how the idea of "distinguishable hypotheses" is important for the VC theorem. If a hypothesis set has finite VC dimension, we can think of it as having finitely many *distinguishable* hypotheses, even if it in fact has infinitely many. Intuitively speaking, if our dataset is "large enough" that not every combination of verdicts is representable with our hypotheses, then we can talk about which hypothesis is truly better than its competitors, as opposed to accidentally matching the specific data points. Having finite VC dimension ensures that there exist finite datasets that are "large enough".

The corresponding concept in model theory relies on the same idea of distinguishability. Intuitively, if a formula is NIP, then there exists a natural number n such that no set larger than that number can be defined using this formula. A model is NIP just in case all of its formulas are (a formal definition is presented in Appendix A; for more details on related concepts, see Simon, 2015).

We can then treat each hypothesis set as a formula defined on some domain. Laskowski (1992) shows that a hypothesis set is VC-learnable just in case the corresponding formula is NIP. What makes this correspondence especially useful is that model theorists have devoted a lot of efforts into determining which model is NIP. Once we know of a model that it's NIP, we also know that any hypothesis sets formulated using the language and domain of this model are VC-learnable.

For example, there is a group of models called *o-minimal*, which roughly means that all the definable subsets of the domain are finite unions of simple topological shapes like intervals and boxes. It suffices for our purposes to note that all o-minimal models are NIP (van den Dries, 1998, p. 90). As it happens, the real numbers with just addition and multiplication are o-minimal (van den Dries, 1998, p. 37). This means that any hypothesis set consisted of addition, multiplication, and the real numbers are going to have finite VC dimension. Similarly, the real numbers with addition, multiplication, and exponentiation is also o-minimal (Wilkie, 1996). This means that all sets of polynomials are VC-learnable, which is a fact already noted by the machine learning community.

As alluded to already, the real numbers with the sine function added are not NIP. This is roughly because, with the sine function, we can define copies of the integers using the set $\{x \in \mathbb{R} : \sin(x) = 0\}$, which allows us to define all of second-order arithmetic, and second-order arithmetic allows coding of arbitrary finite sets. As expected, this is reflected in statistical learning theory by the fact that the set of sine functions has infinite VC dimension, and so is not VC-learnable.

Another important observation from model theoretic investigations on NIP theory is that there seems to be no easy test for when an expansion of the real numbers is NIP. Although the relationship between the NIP property and properties like o-minimal and stable (a set of structures that are not o-minimal but are NIP) is well-researched and understood, there is no uniform way of telling where exactly a model lies (see, e.g., Miller, 2005[3]).

The statistical learning community echoes this difficulty with the observation that "it is not possible to obtain the analytic estimates of the VC dimension in most cases" (Shao et al., 2000; also see Vapnik et al., 1994). Recall that the VC dimension decides how big a dataset is "big enough". If the view is that enumerative induction as a method finds its justified reliability in cases where VC dimension is finite, then our inability to analytically solve the VC dimension of a given hypothesis set seems deeply handicapping.

[3] Technically, Miller is interested in dichotomy theorems which establish either that an expansion of the reals is o-minimal or that it defines second-order arithmetic. As mentioned before, the former suffices for being NIP, and the latter suffices for being not NIP.

To make matters worse, it turns out that even knowing when we do have finite VC dimension is not a straightforward task, as witnessed by the following theorem, whose proof is given in Appendix A

Theorem 1. *The set* $\{\varphi(x,y) : \varphi(x,y) \text{ is } NIP\}$*, where* $\varphi(x,y)$ *is formulated in the language of arithmetic with addition and multiplication, is not decidable. In particular, this set computes* $\emptyset^{(2)}$*, the second Turing jump of the empty set.*

What this theorem tells us is that, in general, there is no effective procedure we can follow that can tell us, for any 2-place formula $\varphi(x,y)$, if it's NIP. With Laskowski's result, this means that we cannot compute, in general, if a given hypothesis set is VC-learnable either.

It is worth noting that this result is about computability *in general* – that is, without reference to the details of the hypotheses involved. There is no effective procedure that decides the VC-learnability of *any arbitrary set of hypotheses.* It is often possible, however, to show the VC-learnability of a particular set of hypotheses on a case-to-case basis. Once we fix an $\mathcal{H}$, we can usually tell if it has finite VC dimension, such as in the cases of polynomials or the sine function. However, this kind of answer is exactly the kind that the discussion of the VC theorem is trying to help avoid. The practical problem of induction can be crudely put as this: there exist inductive methods that sometimes work, sometimes don't, and we would like to know the precise conditions under which something works. The answer from the VC theorem is supposed to be: the method of supervised learning from data works just in case the hypothesis set has a finite VC dimension. However, we now find ourselves in a situation where the condition of having a finite VC dimension sometimes holds, sometimes doesn't, and we do not have a good grasp of the precise circumstances under which it holds. In fact, the above theorem says that we can never have such a grasp.

The specific way in which the set of all NIP formulas is uncomputable is significant also. For some time now, philosophers who study knowledge and learning from a formal perspective have placed a lot of emphasis on learning in the limit. Kelly (1996, p.52), for example, argues that the concept of knowledge (as opposed to, say, mere belief) implies that the method of generating such beliefs is stable in the limit. He then argues that the best way to formalize the notion of "stability in the limit" is to understand it as computable in the limit. Relatedly, a venerable tradition of formal learning theory following Gold (1967) has explored extensively the conditions under which a noncomputable sequence may or may not be approximated by a computable sequence making only finitely many mistakes (cf. Osherson et al., 1986; Jain et al., 1999). From this perspective, it seems we might still be able to claim knowledge of what is or isn't knowable if we can compute the set of NIP formulas in the limit. Unfortunately, this latter task cannot be accomplished. This is because that, in order for a sequence to be approximable in the limit by another sequence, it cannot be harder than the first Turing jump of the sequence used to approximate it (Soare, 1987, p.57; see also Kelly, 1996, p.280). This means that something that is at least as hard as the second Turing jump cannot be approximated by a computable sequence.

To recapitulate the dialectic so far: an easy problem of induction asks us to identify and then justify the conditions under which a given ampliative method is reliable. The VC theorem gives one answer: supervised statistical learning from data is reliable just in case the hypothesis set has finite VC dimension. However, it turns out that we cannot, in general, decide if a hypothesis set is VC-learnable.

3.3 Conclusion

A reasonable conclusion to draw from the discussions we've had so far, I think, is that the VC theorem still does not give us the kind of robust reliability we need to answer a question with some scope of philosophical generality. As is typical of answers people give to problems of induction, as soon as a rule is formulated, a question arises concerning its applicability.

Similarly, what started out as a concern over the robustness of the method of enumerative induction turns into a concern over the robustness of the identifiable (VC-learnable) condition under which enumerative induction is justified to be reliable.

A related question concerns the distinction, if there is one, between the cases where $\mathcal{H}$ has infinite VC dimension and cases where it has a VC dimension so large that it's impractical for us to make use of it. There is a sense in which the case of an infinite VC dimension fails *in principle*, whereas the case of a very large VC dimension only fails in *practice*. However, while there exist ways of empirically estimating the VC dimension of a hypothesis set (see, e.g., Vapnik et al., 1994 and Shao et al., 2000), it is often impossible to analytically solve the VC dimension of a set even if we do know that it's VC-learnable. Together with the result that we cannot test if a case is VC-learnable *in principle*, it seems to suggest that any information we might gain from the distinction between failing in principle and failing in practice will not be very informative, since we often cannot tell which case we are in.

Perhaps the way out is to accept a "piecemeal" solution after all. It seems that when the VC dimension is small, we can often know both that it is finite, and that it is small. And of course how small is "small enough" depends on what the size of a typical dataset is for this particular problem. But again this seems to bring us back to the start of our journey: we can justify the reliability of our preferred inductive method in some cases, but not in other cases, and we cannot offer a unified account on why the good cases are different from the bad ones. Whether this should be seen as a call for further development or a termination of this kind of solution strategy, I leave as a topic for future discussions.

Concluding Remarks

Throughout this dissertation, I have adopted a perhaps unusually strong pragmatic attitude. In each of the three cases examined in the chapters, there exist one or more internally consistent and epistemically sound proposals that prescribe how a piece of methodology is supposed to work. In each case, I raised strictly-pragmatic challenges – there is nothing wrong with the theory; it is simply too practically difficult to get it to work. I have offered some speculation as to how we may respond in ways that both respect practical limitations and keep as much epistemic grounding as possible. I hope that my speculation represents only the beginning of this conversation.

The social sciences are intimately connected to every aspect of our lives. As a result, people's perceptions of them vary wildly from triviality to intractable complexity. While these perceptions may have their merits, it is more productive to start with the assumption that the social sciences sit in between these extremes – they are difficult but ultimately fruitful endeavors. An important part of this assumption is to clarify the logic of social-scientific methodologies not in terms of how they compare with methodologies in physics, but in terms of how they are meant to contribute to our understanding of the social world in a substantive way. This is what I have tried to do in this dissertation.

Another aspect of this assumption is to reflect upon distinctions we draw between social and physical sciences. Throughout this dissertation, I have used the phrase "the social sciences" to loosely refer to sciences that primarily study how people get along socially, such as psychology, sociology, and education. There are, of course, important methodological differences amongst these sciences and important similarities that some of them share with some physical sciences. Indeed, the demarcation between social and physical science is itself a contentious topic in the philosophy of social science. However, I take this question to be only secondary – we need to first clarify what the social sciences are before we can decide whether they differ from other sciences in robust and consistent ways. For this reason, I have remained noncomittal to the demarcation question throughout my discussions.

Whether or not a systematic social versus physical science distinction can be meaningfully drawn, a philosophy of science centered on the social sciences is nevertheless well positioned to contribute to broader philosophical conversations in nuanced ways. For example, although my endorsement of the model-based approach to sampling was motivated by the impossibility of drawing truly random samples of people, the resulting conception of sample representation offers a novel perspective to cases where random sampling is routine. My discussions of measurement validity and data-driven inductive inferences are similarly generalizable to studies of plants, inanimate objects, or particles.

Finally, a philosophy of science centered on the social sciences presents an illuminating perspective on the relationship between science as a knowledge generating endeavor and science as a human activity embedded in a society. With the growing literature on how science affects, and is affected by, its surrounding sociopolitical context, philosophers of science have called for more attention on the interaction between the epistemic and the non-epistemic aspects of science. It is difficult not to see this discourse as an attack on truth, realism, and many other concepts that have traditionally founded their objectivity (and subsequent legitimacy) on a perceived independence from human involvement. Since human involvement lays at the center of the social sciences, any successful philosophy of social science has to contend with this tension in a substantive way, rather than dismissing it as an unfortunate but largely harmless obstacle.

In this dissertation, I have adopted pragmatism as the main approach. That is, I anchor epistemic evaluations of theories to their practical goals and implementations as a way to introduce a kind of pragmatic-epistemic interaction that is neither a strict dichotomy nor “anything goes”. Admittedly, this is not the only possible approach and might just as well not be the best one for resolving problems discussed above. Nevertheless, my hope is that my endeavors in this dissertation will spark more nuanced and sophisticated experimentation in the future.

Bibliography

Abu-Mostafa, Yaser S, Malik Magdon-Ismail, and Hsuan-Tien Lin (2012). *Learning from data*, volume 4. AMLBook Singapore.

Alexandrova, Anna and Daniel M Haybron (2016). Is construct validation valid? *Philosophy of Science*, 83(5): 1098–1109.

Anastasi, Anne (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement*, 10(1): 67–78.

Andrews, Gavin, Scott Henderson, and Wayne Hall (2001). Prevalence, comorbidity, disability and service utilisation: overview of the australian national mental health survey. *The British Journal of Psychiatry*, 178(2): 145–153.

Arnett, Jeffrey J (2008). The neglected 95%: why american psychology needs to become less american. *American Psychologist*, 63(7): 602.

Bingham, Walter V (1937). Aptitudes and aptitude testing.

Boring, Edwin G (1923). Intelligence as the tests test it. *New Republic*, pages 35–37.

Bornstein, Marc H, Justin Jager, and Diane L Putnick (2013). Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Developmental Review*, 33(4): 357–370.

Borsboom, Denny (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Borsboom, Denny (2006). The attack of the psychometricians. *Psychometrika*, 71(3): 425.

Borsboom, Denny, Gideon J Mellenbergh, and Jaap Van Heerden (2003). The theoretical status of latent variables. *Psychological review*, 110(2): 203.

Borsboom, Denny, Gideon J Mellenbergh, and Jaap Van Heerden (2004). The concept of validity. *Psychological review*, 111(4): 1061.

Brewer, KRW (1999). Design-based or prediction-based inference? stratified random vs stratified balanced sampling. *International Statistical Review*, 67(1): 35–47.

Brewer, Ken et al. (2013). Three controversies in the history of survey sampling. *Survey Methodology*, 39(2): 249–262.

Bryson, Maurice C (1976). The literary digest poll: Making of a statistical myth. *The American Statistician*, 30(4): 184–185.

Campbell, Norman R (1938). Symposium: Measurement and its importance for philosophy i. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 17: 121–142.

Cassel, Claes M, Carl E Särndal, and Jan H Wretman (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3): 615–620.

Chang, Hasok (2001). How to take realism beyond foot-stamping. *Philosophy*, 76(295): 5–30.

Chang, Hasok (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.

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

Comaroff, Jean and John L Comaroff (2006). Figuring crime: Quantifacts and the production of the un/real. *Public Culture*, 18(1): 209–246.

Cronbach, Lee J (1980). Validity on parole: How can we go straight? new directions for testing and measure-ment. *Proceedings of the 1979 ETS Invitational Conference*, pages 99–108.

Cronbach, Lee J (1988). Five perspectives on validity argument. *Test validity*, pages 3–17.

Cronbach, Lee J and Paul E Meehl (1955). Construct validity in psychological tests. *Psychological bulletin*, 52(4): 281.

Darrigol, Olivier (2003). Number and measure: Hermann von helmholtz at the crossroads of mathematics, physics, and psychology. *Studies in History and Philosophy of Science Part A*, 34(3): 515–573.

Espeland, Wendy Nelson and Mitchell L Stevens (1998). Commensuration as a social process. *Annual review of sociology*, 24(1): 313–343.

Fiedler, Klaus and Norbert Schwarz (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, 7(1): 45–52.

Gächter, Simon (2010). (dis) advantages of student subjects: what is your research question? *Behavioral and brain sciences*, 33(2-3): 92–93.

Godambe, VP (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2): 269–278.

Godambe, VP (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(2): 310–328.

Gold, E Mark (1967). Language identification in the limit. *Information and control*, 10(5): 447–474.

Graham, Sandra (1992). " most of the subjects were white and middle class": Trends in published research on african americans in selected apa journals, 1970–1989. *American Psychologist*, 47(5): 629.

Guilford, Joy P (1946). New standards for test evaluation. *Educational and psychological measurement*, 6(4): 427–438.

Hall, Christine C Iijima (1997). Cultural malpractice: The growing obsolescence of psychology with the changing us population. *American Psychologist*, 52(6): 642.

Hansen, Morris H, William G Madow, and Benjamin J Tepping (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78(384): 776–793.

Hardcastle, Gary L (1995). S. S. Stevens and the origins of operationism. *Philosophy of Science*, 62(3): 404–424.

Harman, Gilbert and Sanjeev Kulkarni (2012). *Reliable reasoning: Induction and statistical learning theory.* MIT Press.

Hart, Chloe Grace, Aliya Saperstein, Devon Magliozzi, and Laurel Westbrook (2019). Gender and health: Beyond binary categorical measurement. *Journal of health and social behavior*, 60(1): 101–118.

Helmholtz, Hermann von (1887). *Counting and measuring.* D Van Nostrand Company. trans. by Bryan, C., 1930.

Henderson, Scott, Gavin Andrews, and Wayne Hall (2000). Australia's mental health: an overview of the general population survey. *Australian and New Zealand Journal of Psychiatry*, 34(2): 197–205.

Henrich, Joseph, Jean Ensminger, Richard McElreath, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, et al. (2010a). Markets, religion, community size, and the evolution of fairness and punishment. *science*, 327(5972): 1480–1484.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan (2010b). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3): 61–83.

Hoeffding, Wassily (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58: 13–30.

Hölder, O (1901). Die axiome der quantität und die lehre vom mass. *Ber. Verh. Kgl. Sachsis.* trans. by Michell J. and Ernst C., 1996, 1997.

Jacobi, Frank, Hans-Ulrich Wittchen, Christoph Hölting, Sieghard Sommer, Roselind Lieb, Michael Höfler, and Hildegard Pfister (2002). Estimating the prevalence of mental and somatic disorders in the community: aims and methods of the german national health interview and examination survey. *International journal of methods in psychiatric research*, 11(1): 1–18.

Jain, Sanjay, Daniel N Osherson, James Royer, and Arun Sharma (1999). *Systems that learn: an introduction to learning theory*. MIT press.

John, Leslie K, George Loewenstein, and Drazen Prelec (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5): 524–532.

Johnson, Gabbrielle M (2020). Algorithmic bias: on the implicit biases of social technology. *Synthese*, pages 1–21.

Kane, Michael T (2001). Current concerns in validity theory. *Journal of educational Measurement*, 38(4): 319–342.

Kane, Michael T (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1): 1–73.

Kane, Michael T (2013b). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement*, 50(1): 115–122.

Kelly, Kevin T (1996). *The logic of reliable inquiry*. Oxford University Press.

Kessler, Ronald C (1994). The national comorbidity survey of the united states. *International Review of Psychiatry*, 6(4): 365–376.

Kruskal, William and Frederick Mosteller (1980). Representative sampling, iv: The history of the concept in statistics, 1895-1939. *International Statistical Review/Revue Internationale de Statistique*, pages 169–195.

Laskowski, Michael C. (1992). Vapnik-Chervonenkis classes of definable sets. *Journal of the London Mathematical Society*, 45(2): 377–384.

Likert, Rensis (1948). Public opinion polls. *Scientific American*, 179(6): 7–11.

Lin, Zhengyan and Zhidong Bai (2010). *Probability inequalities*. Science Press Beijing, Beijing; Springer, Heidelberg.

Loevinger, Jane (1957). Objective tests as instruments of psychological theory. *Psychological reports*, 3(3): 635–694.

Lowood, Henry E (1990). The calculating forester: quantification, cameral science, and the emergence of scientific forestry management in germany. *The quantifying spirit in the 18th century*, 11: 315–342.

McShane, Blakeley B and Ulf Böckenholt (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, 9(6): 612–625.

Merry, Sally Engle (2016). *The seductions of quantification: Measuring human rights, gender violence, and sex trafficking.* University of Chicago Press.

Messick, Samuel (1989). Validity. In Linn, Robert L, editor, *Educational measurement, 3rd ed.*, chapter 2, pages 13–103. New York: American Council on Education. London: Macmillan Pub. Co.

Michell, Joel (1999). *Measurement in psychology: A critical history of a methodological concept*, volume 53. Cambridge University Press.

Mickelson, Kristin D, Ronald C Kessler, and Phillip R Shaver (1997). Adult attachment in a nationally representative sample. *Journal of personality and social psychology*, 73(5): 1092.

Miller, Chris (2005). Tameness in Expansions of the Real Field. In *Logic Colloquium '01*, volume 20 of *Lecture Notes in Logic*, pages 281–316. Associaton for Symbolic Logic, Urbana, IL.

Neyman, Jerzy (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–625.

Norton, John D (2014). A material dissolution of the problem of induction. *Synthese*, 191(4): 671–690.

Osherson, Daniel N, Michael Stob, and Scott Weinstein (1986). *Systems that learn: An introduction to learning theory for cognitive and computer scientists.* The MIT Press.

Pashler, Harold and Christine R Harris (2012). Is the replicability crisis overblown? three arguments examined. *Perspectives on Psychological Science*, 7(6): 531–536.

Pashler, Harold and Eric-Jan Wagenmakers (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6): 528–530.

Peterson, Robert A (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of consumer research*, 28(3): 450–461.

Pollet, Thomas V and Tamsin K Saxton (2018). How diverse are the samples used in the journals 'evolution & human behavior'and 'evolutionary psychology'? *Evolutionary Psychological Science*, pages 1–12.

Pons, Odile (2013). *Inequalities in analysis and probability.* World Scientific.

Porter, Theodore M (1996). *Trust in numbers: The pursuit of objectivity in science and public life.* Princeton University Press.

Rad, Mostafa Salari, Alison Jane Martingano, and Jeremy Ginges (2018). Toward a psychology of homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, 115(45): 11401–11405.

Rao, JNK and Wayne A Fuller (2017). Sample survey theory and methods: Past, present, and future directions. *Survey Methodology*, 43(2): 145–160.

Royall, Richard (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63(324): 1269–1279.

Royall, Richard M (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2): 377–387.

Royall, Richard M (1992). The model based (prediction) approach to finite population sampling theory. *Lecture Notes-Monograph Series*, 17: 225–240.

Royall, Richard M and Jay Herson (1973). Robust estimation in finite populations. *Journal of the American Statistical Association*, 68(344): 880–893.

Scheaffer, Richard L, William Mendenhall, and Lyman Ott (1971). Elementary survey sampling. belmount.

Scott, James C (1996). State simplifications: nature, space, and people. *Nomos*, 38: 42–85.

Sears, David O (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of personality and social psychology*, 51(3): 515.

Segall, Marshall H, Donald T Campbell, and Melville J Herskovits (1966). The influence of culture on visual perception.

Seng, You Poh (1951). Historical survey of the development of sampling theories and practice. *Journal of the Royal Statistical Society. Series A (General)*, 114(2): 214–231.

Shao, Xuhui, Vladimir Cherkassky, and William Li (2000). Measuring the VC-dimension using optimized experimental design. *Neural computation*, 12(8): 1969–1986.

Shaw, Stuart and Victoria Crisp (2011). Tracing the evolution of validity in educational measurement: Past issues and contemporary challenges. *Research Matters: A Cambridge Assessment Publication*, 11: 14–17.

Shepard, Lorrie A (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2): 5–24.

Shrout, Patrick E and Joseph L Rodgers (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual review of psychology*, 69.

Siegel, Reva B (1994). Home as work: The first woman's rights claims concerning wives' household labor, 1850-1880. *The Yale Law Journal*, 103(5): 1073–1217.

Sifers, Sarah K, Richard W Puddy, Jared S Warren, and Michael C Roberts (2002). Reporting of demographics, methodology, and ethical procedures in journals in pediatric and child psychology. *Journal of Pediatric Psychology*, 27(1): 19–25.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11): 1359–1366.

Simon, Pierre (2015). *A Guide to NIP Theories*. Lecture Notes in Logic. Cambridge University Press, Cambridge.

Simons, Daniel J, Yuichi Shoda, and D Stephen Lindsay (2017). Constraints on generality (cog): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6): 1123–1128.

Smith, TMF (1976). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society: Series A (General)*, 139(2): 183–195.

Smith, TMF (1983). On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society. Series A (General)*, pages 394–403.

Smith, TMF (1991). Post-stratification. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 40(3): 315–323.

Soare, Robert I. (1987). *Recursively Enumerable Sets and Degrees*. Perspectives in Mathematical Logic. Springer, Berlin.

Spearman, Charles E. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, (2): 201–292.

Srivastava, AK (2016). Historical perspective and some recent trends in sample survey applications. *Statistics and Applications*, 14(1-2): 131–143.

Stevens, Stanley Smith (1946). On the theory of scales of measurement. *Science*, 103(2684): 677–680.

Stevens, Stanley Smith (1968). Measurement, statistics, and the schemapiric view. *Science*, 161(3844): 849–856.

Stone, Caroline (2019). A defense and definition of construct validity in psychology. *Philosophy of Science*, 86(5): 1250–1261.

Stroebe, Wolfgang, Tom Postmes, and Russell Spears (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, 7(6): 670–688.

Suppés, Patrick and Joseph L Zinnes (1963). Basic Measurement Theory. In Luce, Bush R. R., R. D. and E. Galanter, editors, *Handbook of Mathematical Psychology, Volume I*, pages 1–76. John Wiley & Sons.

US Census Bureau (1989). *Statistical abstract of the United States, 1989.* Bureau of the Census.

van den Dries, Lou (1998). *Tame Topology and O-Minimal Structures*, volume 248 of *London Mathematical Society Lecture Note Series.* Cambridge University Press, Cambridge.

Vapnik, Vladimir, Esther Levin, and Yann Le Cun (1994). Measuring the VC-dimension of a learning machine. *Neural computation*, 6(5): 851–876.

Westbrook, Laurel and Aliya Saperstein (2015). New categories are not enough: Rethinking the measurement of sex and gender in social surveys. *Gender & Society*, 29(4): 534–560.

Wilkie, A. J. (1996). Model completeness results for expansions of the ordered field of real numbers by restricted pfaffian functions and the exponential function. *Journal of the American Mathematical Society*, 9(4): 1051–1094.

Wintre, Maxine Gallander, Christopher North, and Lorne A Sugar (2001). Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. *Canadian Psychology/Psychologie Canadienne*, 42(3): 216.

Yong, Ed (2012). Bad copy. *Nature*, 485(7398): 298.

Appendix A

Appendix

This appendix presents the proof of Theorem 1 in Chapter Three. I will follow the definition of NIP formulas given by Simon (2015) as follows (with notations changed to match preceding text)

> Let $\varphi(x; y)$ be a partitioned formula. We say that a set A of $|x|$-tuples is *shattered* by $\varphi(x; y)$ if we can find a family $(b_I : I \subseteq A)$ of $|y|$-tuples such that
>
> $$M \models \varphi(a; b_I) \Longleftrightarrow a \in I, \quad \text{for all } a \in A$$

A formula $\varphi(x; y)$ is NIP if no infinite set of $|x|$-tuples is shattered by it.

Following notations from Soare (1987), let W_e to be the domain of the e-th partial recursive function and $Fin = \{e : W_e < \omega\}$.

Lemma Given e, define the following formula in the language of arithmetic

$$\theta_e(x,y) = \exists l > x \; \exists \text{ enumeration } c_1, \ldots c_{2^l}, \text{ first } 2^l \text{ elements of } W_e$$

$$\wedge \; \exists |\sigma| = l \text{ with } y = c_\sigma \wedge \sigma(x) = 1$$

Then $e \in Fin$ iff θ_e is NIP.

Proof. ($\Rightarrow$) Suppose $e \in Fin$. The claim is: there is finite number N such that $|W_e| \leq 2^N$, and for all n, if a set A with cardinality n is shattered by θ_e, then $n \leq N$.

In particular, we show that the claim holds for N being the size of W_e. For the sake of contradiction, suppose there is A, with size n, shattered by θ_e, and $n > N$.

Let $A = \{a_1, \ldots, a_n\}$, $\{b_I : I \subset \{a_1, \ldots, a_n\}\}$, such that $\theta_e(a_i, b_I)$ iff $a_i \in I$.

Without loss of generality, let $a_n \geq n-1$, and $I = \{a_n\}$. Then $a_n \in I$, and $\theta_e(a_n, b_I)$. This means that $\exists l > a_n \geq n-1$ with the first 2^l many elements of W_e enumerated. Recall that the reductio hypothesis states $n > N$. This means that $|W_e| \geq 2^l > 2^{n-1} \geq 2^N$. This contradicts the original assumption that $|W_e| \leq 2^N$.

($\Leftarrow$) To show the contrapositive of this direction, suppose $e \notin Fin$, $|W_e| = \omega$. The claim is: θ_e is IP. Namely, $\forall N \; \exists n \geq N$, with some set A of cardinality n that is shattered by θ_e.

Take an arbitrary $n \geq N$. Let $A = \{0, \ldots, n-1\}$. Let b_σ's be the first 2^n elements of W_e, as σ ranges over finite strings of length n. Since σ is a string, we say $a \in \sigma \Leftrightarrow \sigma(a) = 1$.

We need to show that $\theta_e(a, b_\sigma) \Leftrightarrow \sigma(a) = 1$.

The left to right direction is trivial, since it is part of $\theta_e(a, b_\sigma)$ to state that $\sigma(a) = 1$.

To show the right to left direction, note that since $|W_e| = \omega$, there definitely exists an initial segment of 2^n many elements of W_e, and $n > a$ for all $a \in A$. This satisfies the first conjunct. To satisfy the second conjunct of θ_e, recall that we defined our enumeration to be such that $|\sigma| = n$ with σ being identified with every number $\leq 2^n$. This means that an enumeration of $c_1 \ldots c_{2^n}$ includes all c_σ with $|\sigma| = n$. Define $b_\sigma = c_\sigma$, and we are guaranteed that b_σ is in the enumeration, and $|\sigma| = n$. Finally, the last conjunct of θ_e is satisfied by supposition.

□

Theorem. *The set $\{\varphi(x, y) : \varphi(x, y)$ is $NIP\}$, where $\varphi(x, y)$ is formulated in the language of arithmetic with addition and multiplication, is not decidable. In particular, this set computes $\emptyset^{(2)}$, the second Turing jump of the empty set.*

Proof. Suppose not, then for any formula $\varphi(x, y)$, we can decide if it's NIP. This means that, for any e, we can decide if $\theta_e(x, y)$ as defined in the lemma above is NIP. By lemma, $\theta_e(x, y)$ is NIP just in case $e \in Fin$. If we could decide the former, we would be able to decide the set Fin. But by Soare (1987, p.66, Theorem 3.2), Fin is Σ_2-complete, and so computes $\emptyset^{(2)}$, the second Turing jump of the empty set, and hence is not computable. □