UNIVERSITY OF CALIFORNIA
Santa Barbara

Unveiling Uncertainty: Tools, Concepts, and Philosophy for Defining what is Measured in Educational and Psychological Measurement (and Avoiding a Bewitching)

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Education

by

Daniel Benjamin Katz

Committee in charge:
Professor Andrew Maul, Chair
Professor Diana Arya
Professor Karen Nylund-Gibson

December 2022

The dissertation of Daniel Benjamin Katz is approved.

Diana Arya

Karen Nylund-Gibson

Andrew Maul, Committee Chair

December 2022

ACKNOWLEDGEMENTS

I owe many thanks to many people and I'm sure I could not list them all here. Of course, I owe thanks to my committee. Thanks to Andy Maul whose course sequence in measurement first encouraged me to accept that when I read something that didn't make sense, in fact, it may just not make sense. This, of course, led me down a long, long path of philosophy, statistics, and everything in between. Andy allowed us to pursue what we found interesting and provided assurance when things felt a bit off the wall. To Karen Nylund-Gibson I owe thanks for guidance and instruction in statistical modeling as well as direction in grad school in general. Karen makes GGSE a warm place as a caring mentor. To Diana Arya, beyond just providing mentorship in research direction, writing, and methods, I owe Diana thanks for inviting me into her community of literacy researchers which has kept me motivated in the last few years of grad school. I've enjoyed very much being part of her community of students. Of course there were other faculty members who were a big part of my graduate career such as Mike Wilton who served as unofficial advisor and mentor in navigating research, grad school, and post grad school life and in the end became a wonderful friend. Linda Adler-Kassner who helped me see the pragmatics and necessity of assessment even within a large university (and I always enjoyed speaking with Linda about bikes or other things) and Russ Rumberger who provided much of my hands on training with data analysis, data cleaning, and evaluation.

To the many friends in grad school who made it wonderful, I owe thanks: Melissa Gordon Wolf, for always being up for a debate, coordinating all things social, and really teaching me how to operate at a conference while just being a good friend; Anthony Clairmont, for all the years of collaboration, discussing philosophy and statistics, and getting vegan food; Javier Pulgar for always providing levity, being a great study partner in matrix algebra, multivariate statistics, or really, anything, and inviting me to lots of events(!); Cameron Sublett for talking me into grad school in the first place and discussing things on bike rides; Steve Smith for giving Steve Smith-esque opinions about grad school; and Jay Plasman for always lending a helping hand in various projects and helping me navigate grad school. To students in Andy's and Karen's seminars, thanks for always listening to my long presentations and providing feedback. Especially, thanks to John Galinsky, Jonathan Downey, Adam Garber, and Sabiha Sultana!

Perhaps, more than anyone, thanks to Liz who helped and encouraged me through grad school the entire way, tolerating kindly the far too many hours I put into it while sacrificing many trips and travels. Thanks to the citizens of friends outside of academia who tolerated my 502 ways of being busy. Finally, thanks to my parents and sister, Yael, who urged me to get it done.

Vita of Daniel Benjamin Katz (November 2022)

EDUCATION

University of California, Santa Barbara

Ph.D (2022): Education - Quantitative Research Methods and Philosophy of Measurement
Committee: Andrew Maul (chair), Karen Nylund-Gibson, Diana Arya

M.A. (2017): Education - Quantitative Research Methods
Thesis: Validating a multidimensional measure of reading strategy use

B.A. (2011): Political Science, Minor in History

RESEARCH INTERESTS

Philosophy of Measurement
Philosophy of Language
Literacy
The Rasch Model and Explanatory Item Response Theory Models
Fairness and Ethics in Human Measurement

RELEVANT EXPERIENCE

2022 to Present: NWEA
Role: Research Scientist - Ontology and Measurement

2021 to 2022: Community Based Literacies -
UCSB Role: Measurement Specialist and Analyst
Supervisor: Diana Arya

2021 to January 2022: Microsoft (Vendor/Contractor)
Role: Measure Construction and Psychometrics for User Research

2019-2021: University of Florida, Virtual Learning Lab – IES Grant R305C160004
Role: Graduate Student Psychometrics Extern – Simulation Focus
Supervisor: Anne Corinne Huggins-Manley

2018-2021: UCSB Center for Innovative Teaching, Research, and Learning (CITRAL)
Role: Graduate Student Researcher in Assessment and Evaluation
Supervisor: Linda Adler-Kassner

Summer 2018: New York City Department of Education; Assessment, Design, and Evaluation
Role: Graduate Student Psychometrics Intern – Applying Explanatory IRT Models
Supervisor: Ronli Diakow

2016-2019: California Dropout Research Project (CDRP) Role: Graduate Student Researcher
PI: Russell Rumberger

2016-2018: UCSB Office of Institutional Research
Role: Graduate Student Researcher
PI: Linda Adler-Kassner, Co-Interim Dean of UCSB Undergraduate Education

PUBLISHED WORK

Katz, D, Huggins-Manley, A.C., Leite, W.L. (2022). Personalized Online Learning, Test Fairness, and Educational Measurement: Considering Differential Content Exposure Prior to a High Stakes End-of-Course Exam.*Applied Measurement in Education*.

Arya, D. J., Sultana, S., Levine, S., **Katz, D.**, Galisky, J., Karimi, H. (2022). Raising Critical Readers in the 21st Century: A Case of Assessing Fourth-Grade Reading Abilities and Practices. *Literacy Research: Theory, Method, and Practice*.

Wilton, M, **Katz, D**., Clairmont, A., Gonzalez-Nino, E., Foltz, K., Christoffersen, R., (2021). Improving academic performance and retention of first-year biology students through a scalable peer-mentorship program. *CBE - Life Sciences Education*.

Arya, Diana, Clairmont, A. **Katz, D** & Maul, A. (2020) Measuring Reading Strategy Use, *Educational Assessment*, 25:1, 5-30, DOI: 10.1080/10627197.2019.1702464

Maul, A and **Katz, D**. (2018). Internal Validity. In B. Frey (Editor), The SAGE Ency- clopedia of Educational Research, Measurement, and Evaluation. Thousand Oaks, CA: SAGE.

Katz, D. (2017, May). An Update: The Narrowing California High School Graduation Gap between Black, Latino, and White Students. (CDRP Brief No. 24). http://cdrpsb.org/pubs_statbriefs.htm

Katz, D. (2017, March). The Narrowing California High School Graduation Gap between Black, Latino, and White Students. (CDRP Statistical Brief No. 23). Retrieved from http://cdrpsb.org/pubs_statbriefs.htm. The Narrowing California High School Graduation Gap between Black, Latino, and White Students

CONFERENCE PRESENTATIONS

Clairmont, A. & **Katz, D.** (April, 2021). Using Rasch Measurement Theory for Responsive Program Evaluation. Paper presented at the American Educational Research Association Annual Meeting (AERA). To be held Virtually

Katz, D, Huggins-Manley, A. C., & Leite, Walter (July, 2021). Technology-Enhanced Learning Platforms, Opportunity to Learn, and Test Fairness. Paper to be presented at the International Meeting of the Psychometric Society (IMPS).

Katz, D and Clairmont A. (April 2020). Should Psychometricians Make Claims About Test Fairness? Paper to be presented at the National Council on Measurement in Education (NCME) (Postponed due to COVID).

Katz, D. and Diakow, R. (April, 2019). Using Explanatory Item Response Theory Models to Re-Examine Fairness in Psychometrics. Paper to be presented at the National Council on Measurement in Education (NCME) April 4-8, 2019. Toronto.

Katz, D., Nylund-Gibson, K., & Furlong, M. (2019, April). Is One Item Enough? Examining Affect-Laden Survey Items Using Mixture Modelling with Distal Outcomes. Paper presented at American Educational Research Association Annual Meeting, April 5-9, 2019. Toronto.
Best Graduate Student Paper in the AERA Survey Special Interest Group Sig

Katz, D., Clairmont, A., Arya, D., & Maul, A., (2018, April). Measuring Reading Strategy Use in a Multilingual Context. Paper presented at American Educational Research Association Annual Meeting, April 13-17, 2018. New York.

Katz, D., Clairmont, A., Arya, D. & Maul, A. (2018, April). Measuring Reading Strategy Use in a Multidimensional, Multilingual Context. Paper presented at the International Objective Measurement Workshop (IOMW), April 10-12, 2018. New York.

SELECT IN PROGRESSWORK

Katz, D., Maul A., Clairmont, A (2021) Tools from philosophy of language for increasing transparency in psychological measurement”

Clairmont, A. & **Katz D.** (in preperation) Using Rasch Measurement Theory for program evaluation: A methods note. To be submitted to *American Journal of Evaluation*

Katz D. Reconsidering fairness in educational assessment: using tools from ethics, causal inference, and psychometrics

TEACHING EXPERIENCE

Instructor

One Day Workshops
UCSB Methods U - An Introduction to R: Data Cleaning, Wrangling, and Visualizing

Implementing Explanatory IRT via the R package, TAM

Teaching Assistant

Fall 2020: Education 217C, Constructing Measures
Spring 2018: Education 214C, Linear Models for Data Analysis (and categorical data analysis)
Winter 2018: Education 214B,Inferential Statistics
Fall 2017: Education 214A, Introductory Statistics

ORGANIZATION & COMMITTEE MEMBERSHIPS

Graduate Student Representative - NCME Committee on Informing Assessment Policy and Practice
American Educational Research
Association National Council on
Measurement in Education Society for
the Improvement of Psychological
Sciences

STATISTICAL SOFTWARE

R
Commonly used stats packages: TAM, mirt, lme4, Lavaan, plm
Stan, brms, rstan (Basic Knowledge)
Other packages: tidyverse packages, Rmarkdown xaringan
Mplus
Stata (Basic Knowledge)
Tableau (Basic Knowledge)
Git and Gitub

ABSTRACT

Unveiling Uncertainty: Tools, Concepts, and Philosophy for Defining what is Measured in Educational and Psychological Measurement (and Avoiding a Bewitching)

Daniel Benjamin Katz

The quality of a measuring instrument is always to some extent limited by the conceptual clarity and coherence of the target property's (that which we would like to measure) definition. However, foundational texts about validity theories, psychometrics, and latent variable modeling offer little (non-operationalist) guidance for transparently defining or analyzing these properties (sometimes called, constructs, attributes, factors, latent variables, or similar). I begin by suggesting, contrary to many methodological approaches, that non-operationally defining properties is the first necessary step in any measurement effort and an ethical imperative in the human sciences. Productive definitional efforts are those that enable what Hasok Chang calls epistemic iteration. This means that definitional efforts serve certain initial scientific aims but should be ever improved and critiqued. I try to present useful recommendations for productive definitional work from the fields of psychometrics, philosophies of language and science, as well as metrology and exploratory statistics. That we know enough about a property to define it and treat it as measurable is presented as an empirical and ethical question.

Throughout, I use running examples from definitional debates surrounding the assessment of reading comprehension ability in the National Assessment of Educational

Progress (NAEP) and scholarship related to “non-academic” student attributes such as human resilience. In the final section of this dissertation, I present an empirical example from a reading test, repurposing differential item functioning (DIF) and measurement invariance testing from psychometrics as useful inductive tools for demarcating properties we would like to measure. Ideally, these methods for property definition can help facilitate conceptual and linguistic clarity for measurement in education, psychology, and the human sciences more broadly.

Table of Contents

Chapter 1
Introduction to the problem of definition in psychometrics

In his paper, "Attack of the psychometricians," Denny Borsboom argues that in an ideal world:

> "The first thing a psychologist who has proposed a measure for a theoretical attribute would do is to spell out the nature and form of the relationship between the attribute and its putative measures." (Borsboom, 2006, p.429)

Borsboom complains, of course, this is not what happens. Typically, he says, in the setting where a researcher wishes to measure something, the process involves first selecting among the quiver of typical statistical models and then fitting those models to data. What lacks from this mode of research is an account of the link between the way the results of a test or survey came to be and the theory of the attribute of interest. Borsboom claims this disconnect between what is being measured and the results of measurement process makes it is unclear which hypotheses are being posited and tested exactly or even theoretically and conceptually unassailable because nothing is specific enough to critique or improve. Borsboom charges psychometricians with perpetrating this improper thinking because of an overcommitment to the logical empiricist roots of classical test theory (CTT) – a devotion to meaningful research in terms of only the empirical or so-called observational. He also argues that there is an atheoretical devotion to patterns in the data without consideration of the data generating process – which might require an invocation of theoretical entities.

Haig & Borsboom (2008) point out bluntly, “for a procedure to count as a measurement procedure, it must yield measurements of something; that is, it requires that there be a certain connection between the observations and some theoretical attribute” (p. 2). The term *theoretical attribute* may need some clarification in usage for it to make sense in the context of measurement. For instance, many probably would not consider a person’s height to be theoretical, though measuring a person’s *true* height might be considered sort of hypothetical. Nonetheless, the point is well taken – measurement requires delimiting what is being measured, what is not being measured, and what might be influencing measurement results. So how would one go about defining the property one would like to measure? What is this process like for scientific endeavors? Where can one look for guidance in this area?

Hibberd (2019) laments, in fact, that *scientific definition*, especially for the purpose of measurement, is not actually covered in psychological methods textbooks. Hibberd seems to use the term *scientific* to delineate definition that might be used for scientific inquiry which aim to refer to existing properties (property is a term discussed in chapter 3 of this dissertation) from commonly taught *operational* definition. Like Borsboom, Hibberd decries the over commitment to logical empiricism and homes in on operationalism. Operational definition, from operationalism, is the process of defining something by the method used to measure or assess it. An example of operationalism would be constructing two different assessments of reading comprehension ability that each comprises 5 test item, or test questions – really any form of observation $\{x_i \ldots x_n\}$): $X_A = \{x_1, x_2, x_3, x_4, x_5\}$ and $X_B = \{x_5, x_6, x_7, x_8, x_9, x_{10}\}$. The operationalist would say that there are now two different *reading comprehension abilities* – scores from *reading comprehension abilities* $X_A = \{x_1, x_2, x_3, x_4, x_5\}$ and *reading comprehension ability* $X_B = \{x_5, x_6, x_7, x_8, x_9, x_{10}\}$. Each time a new test is constructed, one would be creating a new reading comprehension ability. This is different than just saying that test X_A and test X_B are sensitive to different influences such that results are not comparable without any form of statistical adjustment. Note, the item x_5 appears in both tests. In psychometrics, a common item might be used to place the tests scores on a common scale (e.g. Kolen & Brennan, 2014). Alternatively, we might use a computer adaptive test (CAT), to administer fewer items to students so that few students see the exact same items. These scores are only interpretable and comparable if we do not take an operationalist position because different observations are used to make inferences about the same thing (e.g., reading comprehension ability).

The physicist Percy Bridgman, the first explicator of operationalism, said "we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations" (Bridgman, 1927, p.5). Bridgman further required that "the set of operations equivalent to any concept be a unique set, for otherwise there are possibilities of ambiguity in practical applications" (Bridgman, 1927, p.6). In other words, concepts each required a specific operation. Many scholars have pointed out, including Bridgman himself and Hibberd, that operationalism is problematic as a foundation for scientific inquiry because it confuses the property under measurement for the way one attempts to measure it. Weird problems arise with operationalism such as lengths measured via two different methods leading to two different length concepts. So operationalism does not provide a way out for scientific definition. However, the operationalist position, which provides us with a lack of uncertainty about what comprises the property that is measured, when turned into a realist position about what is measured, is turned into uncertainty in the form of now having imperfect definitions of what is measured in, say, a CAT test, since one now has the notion of a “purified” property that must be demarcated from other properties (e.g. reading comprehension ability demarcated from background knowledge about a textual area). However, it is not often known how to manifest this property in any idealized form. To make it clearer, imagine, the *true height* of Haig and Borsboom (2008). The true height of a person is unknowable and idealized since there are always perturbations in measurement processes that may come from slightly different conditions of measurement each time (especially if different instruments are used) and the very definition of what and where one should measure (for instance, we might say the best measurement of the height of person *x* should be taken while they are wearing no shoes and standing straight up. Thus, the height will then be changed by how straight the person stands).

A cross cutting theme between Hibberd and Borsboom is that definitions for scientific or knowledge gaining purposes are not just arbitrary but imply that there is something to be identified by a term or word used – the term “reading ability” supposedly demarcates something in the world, it is not just a free-floating entity that can take on any arbitrary meaning. What a term refers to is also not an arbitrary combination of entities[1] (Meehl, 1992 provides an example of arbitrary groupings – for our own example, we can imagine grad students who study educational measurement but also ride bikes; or for a measurement purpose, from Mari, Wilson, & Maul, 2021 – the measurable properties of height and weight could be combined in some way). Borsboom and Hibberd call for a scientifically realist account of what one aims to measure– definitions designate something in the world. This

[1] Later, including in chapters 2 & 3, I’ll attempt to introduce these concepts as *universals*. Universals are defined as something like, mind/subject independent, real or existing, and nonarbitrary. For instance, the property of *length* is a property of an object like a dog, and there are many instantiations (realizations) that height can take. Length values for all the things that have lengths have in common the property of length.

designation, though, in the case of research, may have both ontological and epistemological implications. The language has ontological implications because it is a commentary on what there is, and epistemological implications because the language directs investigation.

Despite problems discussed above, operationalist language is not uncommon in psychology makes its way into textbooks. For instance, in the oft-used textbook *Confirmatory Factor Analysis for Applied Research*, which has no fewer than 21,500 citations according to Google Scholar, the author provides an example of so-called construct validation by saying that “juvenile delinquency may be construed as a multidimensional construct defined by various forms of misconduct (e.g., property crimes, interpersonal violence, participation in drugs, academic misconduct)” (Brown, 2015, p. 2). Since this book is about confirmatory factor analysis (CFA), a reflective measurement model in which factors or “constructs” are meant to cause the item responses might be assumed, but the interpretation of juvenile delinquency becomes difficult. The definition is formative in the sense that juvenile delinquency is arbitrarily defined by the authors. Is juvenile delinquency caused in the sense that the four factors together cause scores on a higher order or factor (see, for instance, Edwards & Bagozzi, 2000) Is something being constructed in the sense that juvenile delinquency is merely a label for a convenient, however defined, set of behaviors deemed problematic but is not necessarily a single property of persons? A consequence of this operationalism is that there is no way to improve the measurement or theorizing around whatever juvenile delinquency is (is it a mental state? An attribute of a person? Only a set of actions?). The score from an assessment of juvenile delinquency is, borrowing language from Edwards and Bagozzi (2000), juvenile delinquency itself. An even more egregious example from Knight et al., (2009), also a textbook, says, “configural invariance is established if a

CFA model that allows the same set of items to form a factor in each group shows good model fit" (p. 109). Clearly, this is problematic since, in a confirmatory factor model as expressed in the text, item responses are supposed to be caused by their factors and do not define, or *form*, as the authors say, a factor. Again, the problem is that factors in a reflective CFA model are supposed to have existences independent of how they are measured. Borsboom et al. (2003) argue that making meaning of CFA models or other reflective measurement models requires a scientific realist orientation. It is argued next that not taking some sort of stance about definition can present real world scientific and ethical challenges.

CFA models in psychometrics, are special cases of a general class of models called latent variable models. It is not uncommon for researchers in psychological measurement to say that what they would like to measure is a latent variable. The basic notion, as discussed later in this dissertation, is that there is something unobserved about a particular property that is often called a "variable", and the unobservability is attributed to a suite of reasons – for instance, that the variable is merely hypothetical, the variable is theoretical, constructed, a latent factor, or that unobservability is simply a feature of the property itself. In other words, something like "reading comprehension ability" is in principle unobservable (ignoring for now, the term *variable*, a mathematical entity, used in place of the term *property*). Borsboom (2008) provides an account of why the distinction between unobservable and observable (or latent and observed) is a false dichotomy. Maxwell (2009) provides an account of, why in general, in many cases, the observable-unobservable distinction in general is also likely a false, if unknowable, distinction to make as a characteristic property of entities. Both authors appeal to the notion of, among other points, the improved instrumentation and research base as science progresses that may advance something unobserved to be treated as observed.

However, for the most part, neither author designates the consensus about the property of interest as being a source of what might orient our classification about the present observability of a phenomenon. Both authors do note that the observability and unobservability phenomenon is a *human* distinction and it would be hard to say that unobservability is a characteristic of the properties we'd like to measure. This argument will be discussed further in chapter 2, but I would like to posit that part of what gives us the account of something being observable (or not) is a matter of human consensus about the property under measurement and trust in theoretical background assumptions. Classifying something then as unobserved (or unobservable) or observed (or observable) is, at least in part, about consensus around its defining features. In this way, it can be studied (else, how would we demarcate what is to be studied) by multiple research groups, begin to figure in theories or experiments with explanatory roles and interpreted in the same way across contexts. This makes defining something as a latent variable, or as latent, a move to either construct an unassailable black box (how can you ever debate what something is if it is perennially unobserved?) or an admission of high uncertainty.

Some have argued that part of the reproducibility problem can be attributed to measurement problems (e.g. Flake, 2021; Loken & Gelman, 2017). While much of the reproducibility world has been devoted to increasing transparency and openness of analytic decisions (e.g. posting code, data), it is less clear how to be transparent about the definition of what is measured. Or, in other words, it is less clear about how to define what we would like to measure and how those definitions might be integrated into analytic decisions. At least part of the open science and reproducibility realm is about building consensus. When claims about certain psychological entities are made based on measurements of those entities, consensus is necessary for speaking to each other or interpreting results. This requires both semantic (devotion to the meanings of terms or parameters in models) and more technical and syntactic (structure of models, statistical or otherwise) consensus. There is considerably little work in how to start building this consensus in education or psychology. I maintain that a devotion to latent variables or attributing to things that we'd like to measure as latent being part of its definition not one of those methods. This is not to say that latent variable statistical models should not be used. But, separating the statistical commitments in a model to the substantive commitments is a necessary first step. This lack of consideration can lead to confusion and unproductive conversation. I aim to provide some tooling for making conversations about what is measured more transparent and productive. Operationalism may, to some degree, be maximally transparent from the perspective of what a property label means, but it is not fallible because it cannot misrepresent. If you subscribe to any sort of scientific realism (that there are mind independent entities, the goals of science are to describe and explain these entities, and our knowledge of these entities is ever being improved), then we cannot use operationalism since it does not matter whether there is some matching between the way the world is and the measurement procedure. Alternatively, at the far end of latent variable theory, we can be tempted into describing an attribute as permanently ethereal. I aim to build on these views using some examples from education and psychology.

1.1 Characterizing Descriptions of Measurement

Given the discussion above, a definition of measurement is necessary. Drawing on writings from researchers in measurement (e.g. Giordani & Mari, 2012; Mari, 2013; Mari et al., 2012a, 2019), I shall try to lay out some basic tenets of what measurement is and what it is not. First, I take measurement to be a human process created by humans for a purpose. As such, this means the definition can always change for a given purpose. Definitions of measurement largely track with what has been considered measurement, and trusted measurement, now and in the past. Measurement is indeed an “evaluation process” (Mari et al., 2012, p. 2108). The descriptions of measurement in this section are not meant to be definitive so much as to lay the foundations of what definition of what is measured for the purpose of measurement might look like given what measurement requires. For our sake, what we consider measurement leads to measurement results that are trusted and can be interpreted across time and place. Second, measurement in psychology, also, did not originate on its own, but was rather brought about by physicists or biologists trying to measure things like human sense perception or intelligence (e.g. Michell, 1999; Mulaik, 1985; Tal, 2019). As such, the question about measurement’s possibility in psychology and education is answerable to the extent that there is some similarity correspondence to measurement in the physical and natural sciences since this is where the very idea came from. Therefore, definitions of measurement should be largely drawn from measurement’s success in physical sciences. Given these two features, a one sentence, pithy definition just is not possible.

The first part of this characterization will involve the outcomes of measurement. What has been considered measurement leads to trustworthy results from a measurement process. That is, measurement leads to results that can lead to actions – for instance, fitting a train through a tunnel. However, for the measurement to be trustworthy and actionable, results must be interpretable across time and people. That is what Mari et al., (2017) call intersubjectivity. For the measurement result to be intersubjective, results must be known to be about or in reference to the same property across contexts. The target property under measurement will be called the measurand. Mari et. al (2012) call this objectivity. That is, we can get the train through the tunnel because the measurements associated with the height and width of the tunnel and the train can be interpreted in the same way by different engineers working on different projects. It is known that these height and width measurements, likely coming with some uncertainty quantification, can be primarily attributed to the length or width in given dimensions, making, e.g., *width of the tunnel at its narrowest point* .the measurand. This requires measurement units and requires output data that is evaluated by people. This may eliminate certain uses of latent variable models (e.g. factor analysis, IRT) as measurement since there is no easily interpretable unit or outcome (e.g. factor/scale indeterminacy problems) that can be interpreted across time and place. Without a measurand, it is somewhat hard to even make a judgement about how much trust should be attributed to measurement results (or uncertainty attributed to the measurement results). This places the property under measurement and ideas about the way it exists, as well as the notion of consensus, front and center.

1.1.1 A Model of Data Generation

The second part of the characterization here will be about how the results come to be – how they come to be trusted. This portrays measurement as a process. Measurement is cast by Mari (2012) as solving a measurement problem in order to evaluate and provide information about a property of something. An object or entity has certain properties (characteristics). These characteristics are things like the height of an object or weight of an object where height and weight are properties. Thus, height and weight must be uniquely identified and defined in order for the measuring instrument to interact with the properties. The general process of measurement will involve putting a measurement instrument in interaction with an object and the measuring instrument (e.g. ruler, thermometer, test) indication changes because of the way the property of interest interacts with the instrument. An example would be gas expanding in a tube when a thermometer is put in contact with an object that is warmer than room temperature, say. This requires a theory of how the property of interest, the thermometer, and other properties interact (e.g. Tal, 2016, 2019). This leads to an instrument indication that might be placed on a scale (e.g., expansion of gas in a tube is an indication that is placed on the Celsius temperature scale). A fuller description of the process is given by (Mari et al., 2021). Again, this places the property to be measured as the most important thing in the measurement process. However, this also implies that measurement is not just about estimation of values. Though, that may be a *part* of measurement (for instance, estimating ability from raw test question responses), measurement is also characterized by the entire process and a definition of measurement could not simply be about the assignment of numbers (discussed further below; Stevens, 1946). Measurement is then model reliant – because the properties under measurement will be idealized and unobserved since there will always be some error. Therefore, a model is required which idealized the measurement system is required – the process of measurement, the object under measurement, and the property itself all need to be modeled. This model gives information about what values can be reasonably attributed to an object. The field of metrology is largely devoted to defining and specifying what goes into measurement – but has been primarily concerned with measurement in the physical sciences. It has little concern for the definition of properties that are less well defined. Saying that, given this measurement process, the field of metrology given its long history (discussed more in chapter 2) compared to psychological measurement, provides a useful framework for defining and characterizing measurement. Discussions of whether measurement entails only quantities or can also include classification as part of measurement may be beyond the scope of this dissertation. However, for now, I will say that measurement involves at least working with continuous quantities and ordinal attributes but may exclude classificatory activities. The field of metrology has a definition of measurement that leaves room for interpretation but reads:

> process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity (Joint Committee for Guides in Metrology (JCGM), 2012, 2.1);

Note, this differs from traditional definitions of measurement in psychology attributed to Stevens (1946; “the assignment of numbers according to rules”). When combined, these two areas - the process and how it leads to an outcome - characterizes measurement as a human evaluation of a property of interest leading to empirically derived (at least in part) values that can be attributed to the property in that case. The reasonableness of attribution of values places knowledge of the property at the forefront and hence, a clear definition is important. Given that we are using terms like “interact” or “select”, measurement is thus, in some capacity, a causal process. Changes in the property of interest cause a different transduction and interpretation of the instrument indication.

What does this look like in a test or survey? A student takes a test where the test is a measuring instrument (though it need not be). The test is designed so as when a student responds to questions, the source of the response comes primarily from the property of interest. That means items or test questions or observations are, in some capacity, interacting with a student’s cognition. Surely, there are other properties that affect student responses and these might be called influence quantities as in the International Vocabulary of Metrology (VIM). Foregrounding the discussion in chapter 4, psychometric language and tradition has developed its own weaker vocabulary (weaker as in less committed to causal language) such as referring to influence quantities as “construct irrelevant variance” (Messick, 1995). These are things that might change the reading of an instrument – for instance, the text complexity of a word problem in a math test might change a student’s response but does not change their math ability. However, in some cases, we may be able to model this system. Chapter’s 4 of this dissertation will be particularly interested in these influence quantities and how they may actually help us think about the properties of interest. It is contended that the lack of semantic clarity relating to how to speak of definitions of properties we intend to measure, sources of uncertainty, and the language used to describe the properties often leads to unproductive conversation. This is not to say the language of psychometrics has not been useful so much as to say there is no reason to stop at the language of psychometrics.

1.1.2 Why worry about what measurement is?

Why worry about characterizations of measurement? One (1) perspective is simply historical. This is the continuation of efforts from the likes of Fechner who was so committed to applying the method of physics to the version of quantitative psychology that he created became known as psychophysics – therefore, this effort to define measurement is part of the continued history (Fechner, 1987; Michell, 1999). Another related perspective (2) is the quantitative imperative belief imported from the likes of Kelvin and transferred to early researchers in psychology like Cattell, Boring, Spearman (Michell, 2003), and perhaps to people like Quetelet and Galton (e.g. Mulaik, 1985; and summarized by Michell, 2003). This is the idea that if something cannot be quantified or measured, then we do not (or cannot) truly know it. Thus, position (2) means that for psychology to be taken seriously as a science but also to have command over its content, it must be quantifying. A third (3) position might be to discuss knowledge sharing improvements and enquiry. Thus, it is carving out what is relevant in a discussion and what is not when discussing measurability or measurement in general of psychological phenomena. Consider the discussion about uses of latent variable models that do not lead to an indication about the value of a property under measurement. This should lead us to ask about what is being answered with the use of these latent variable models – what are they accomplishing for us? A fourth (4) position could be something along the lines of what the field of metrology is trying to accomplish – to set guidelines for improving and admitting certain evaluations of the world that take a certain form and assess their trustworthiness or alignment. This is about measuring for purpose. Since measurement is often associated with trustworthiness of information, we might ask, if we are to measure, what would lead to certain results being trustworthy? I am worrying throughout this dissertation about what characterizes measurement as a cross between points (3) and (4). Many forms of research in education and psychology are treated or claimed to be measurements with associated instruments we might call measures, or less elegantly or accurately, "scales." Should claims from these instruments (surveys, academic tests)mn be afforded the same trustworthiness we might afford measures of the physical realm? If they are, at least in part, given the characterization above, we should have some concept of what is indeed measured and modeled.

1.2 Reading, Resilience and Definitional Uncertainty

Consider the so-called Nation's Report Card - the National Assessment of Educational Progress (NAEP) which is a congressionally mandated testing program in the

United States. It assesses 4th, 8th, and 12th grade students in the areas of math, reading, and science (among others), providing information at the national, state, district, and subgroup level but not the individual student level. Meant to be a policy lever (more on this below and in chapter 3), NAEP is said to provide a "common measure of student achievement across the country" (U.S. Department of Education, 2010, p. 3) . The National Assessment Governing Board (NAGB) creates frameworks for assessment areas describing what students should "know and be able to do" (U.S. Department of Education, 2010, p.4) and describes the sorts of content that should appear in NAEP tests as well as how scores should be reported. The NAEP document quoted above does not say this, but content and the frameworks about what students should be able to know and do are inherently related, of course. The 2025/26 proposed framework update is the first real update to the NAEP reading framework since 2009 .and is hotly debated.

What is the source of contention? Some scholars and state officials have wanted to revamp NAEP via a so-called new definition of reading comprehension, though, most debates surround what content should be on the NAEP assessment. For instance, should NAEP reading items require certain forms of background knowledge or should that background knowledge be supplied in the test itself? This is a test content question that rests on how one might define, "reading comprehension" and what one may consider a relevant observation. However, one worry is that since NAEP is used for tracking trends over time, changing content or the definition of what is "measured"[2] will mean that reading scores will not be comparable. Note, the operationalist discussion above. If one were to take the operationalist position, any change in content would yield incomparable scores *regardless* of

[2] I use quotes here because I'm not sure if the NAEP is worried about measurement or testing and assessment.

statistical adjustments, because the method of measuring has changed and hence a different reading comprehension is measured. However, if one were to take a realist stance, if the debate is about changing content to better measure reading comprehension, it may (though, no guarantee) be possible to compare scores over time, but different measurement and statistical tools would be needed. However, if the very thing in the world that is being measured is changing, then scores are definitionally incomparable (e.g. Nitko, 2016; Tal, 2019; one cannot say that 3 inches is the same `amount` as 3 pounds – this is incoherent).

With NAEP, it is not quite clear what this intention is. For instance, it states that the new framework "is updated to reflect three research-based developments that help ensure that the NAEP Reading Assessment remains a useful measure of reading comprehension"(National Assessment Governing Board, 2021, p. 13). This would seem to be the position that measurement of the same property is being improved but what is being measured is not changing. In other words, is this about changing content of the test to better reflect a current definition of reading comprehension or is it changing the definition of reading comprehension, or improving it, even, to better reflect what current researchers believe about what the property of *reading comprehension* is. The NAEP governing board separately set out as a goal, to "expand the construct of reading" (National Assessment Governing Board, 2021, p. 3). It is not clear exactly what this latter quote means either, but it seems that it could be *potentially* inconsistent with the first quote in this paragraph. Depending on the position of what a construct is (a property to be measured or something more operationalist), it seems for any results from NAEP to be interpretable, expansion of the definition of reading comprehension is not *necessarily* the same as just improving its measurement instrumentally – improving the sorts of items and observations used - but improving its measurement by *updating* what reading comprehension ability as a property of people refers to (a possibility is that we also have better knowledge of reading comprehension) which also demands different items and observational settings. It is clear some thinking about *why* these are stated aims is important since, theoretically, policy and curricula may be altered based on the words used to describe this improvement.

Sorting out the different options, I think, will help clarify the goals of something like NAEP reading but also better frame debates around it. A coherent, or at least transparent presentation of terms would be useful – however, there are few tools for being transparent definitionally. It seems, based on articles in press outlets, education blogs, and to some extent, academic journal articles, there have emerged two sides of the NAEP debate. One side argues that integrating background knowledge into texts students have to read is an important part of reading comprehension but for NAEP, the background knowledge is not explicitly part of the main attribute of interest - reading comprehension - which is supposed to be a cognitive process. Hence, key terms or knowledge required to answer a question that might be obscure to some respondents, should be handled through providing students' explicit definitions of these terms, for instance. The other side of the debate seems to argue that, reading requires making sense of background knowledge and words one has never seen, so providing knowledge scaffolds is problematic for score interpretation (e.g. Finn, 2021; NAGB, 2021; Schwartz, 2021). There is also worry that changing too much will make scores across time incomparable.

Using NAEP reading as a running example, I hope to show that tools for paying attention to the language about what is measured and what measurement is, is not only useful for these debates in reading, but is also an ethical imperative. Framing two sides of the debate about what should be included in NAEP reading will provide a base for showing the relevance for taking definition in measurement seriously. Thinking about literacy in general Scribner (1984) argues that definitional debates have "more than academic significance. Each formulation to an answer to the question, "What is literacy?" leads to a different evaluation of the scope of the problem (i.e. the extent of illiteracy) and to different objectives for programs aimed at the formation of a literate citizenry" (Scribner, 1984, p. 6). Scribner (1984) criticizes certain definitional efforts as trying to find the best definition of literacy as a property of individuals. However, she argues, this is a problem because "literacy is a *social* achievement . . . Literacy is an outcome of cultural transmission…it follows that individual literacy is relative to social literacy" (p. 8). This implies that defining literacy, of which we might consider *reading comprehension* embedded within, requires considering what *presently* counts as literacy. That is, in the context of literacy, we must confront different concepts of "value, philosophy, and ideology" (Scribner, p. 8) when it comes to defining the social act that is literacy. Later in this dissertation, we will note that admitting certain psychological attributes as real (or into our ontologies as something that can be studied) requires admitting that social construction does not render something unreal. Clearly, from Scribner, we cannot admit these things without questions such as, "why do we study this? What about "this" makes it interesting? Where did the notion of the thing we would like to study come from and who does it benefit?" Literacy is one example. Paying attention to the structure of debates is important. However, on the psychometric side, there is little writing or direction about how to do this. One way to get a bird's eye view, though, might be to compare methods used in different areas.

In other areas of education policy, but also relevant to psychological research and practice, there has been an emphasis on measuring and reporting student attributes not traditionally associated with academic skills (academic skills such as a reading and writing and doing math). A popular interest in social emotional learning (SEL) has taken root as a policy initiative for reporting (for instance, the PACE CORE schools initiative in California, which, among other things, has focused on measuring and increasing certain "constructs" such as growth-mindset, self-efficacy, and social awareness; see, for example, Gehlbach & Hough, 2018; Marsh et al., 2018). Notably, the definitional boundaries, or what we know about the distinguishability of different properties are slim, if blurry. Marsh et al. (2018) say that one reason for a lack of implementation of SEL in schools is because "the definition of SEL and what constitutes high-quality SEL support and instruction are often elusive and unclear" (p. 4). Perhaps in contradiction, the authors also cite studies supporting the claim that certain aspects of student SEL have effects on student outcomes leading one to wonder if we don't know what a term like "self-efficacy" refers to, how we would know that the phenomenon of interest is indeed what caused positive effect. Relevant to both teachers and studies in psychology, the term "resilience" is often used to describe a property of people (though, in some definitional efforts, these are not just limited to the human but the situation the human is in).

Scholarship on human psychological resilience research has important, altruistic goals focused on figuring out why some people overcome various challenges and others do not. In early resilience research, those who overcome challenges are labeled the resilient ones (more in Chapter 3). Being able to track down why and how somebody was resilient and somebody else was not has intuitive benefits for those working in fields such as counseling or clinical psychology or even teaching. The rough idea is that resilience, conceptualized as a nebulous composite concept that encapsulates a number of phenomena, is a psychological *construct* that causes some to overcome challenges and others to falter. Naturally, there has been some desire to measure "it" (Connor & Davidson, 2003; Wagnild & Young, 1993; Windle et al., 2011). Yet, there are unsettled definitional disputes, that demarcate, nearly completely different aspects of reality. For instance, in a noble effort, Southwick, Masten, Panter-Brick, and Yehuda (2014a) jointly wrote a paper consisting of a series of responses to each other about resilience, focusing on, among other things, several definitions of resilience. However, this dispute seems intractable for the wrong reasons. While reading comprehension in NAEP debates are to some extent about values or admitting some activity or process into the definition of reading comprehension, works like Southwick et. al. (2014) seem to be primarily semantic and lexical. For instance, in the panel discussion, Southwick, argues that the American Psychological Associations definition of resilience at the time (2014):

> ‘‘the process of adapting well in the face of adversity, trauma, tragedy, threats or even significant sources of stress (para. 4).’’ (Southwick et. al. 2014, p.2)

is not adequate because it "does not reflect the complex nature of resilience" and that "determinants of resilience include a host of biological, psychological, social, and cultural factors" (both quotes in Southwick et al., 2014b, p. 2). Note what's being asked of this definition. First, it is not clear how causes of resilience belong in a definition of resilience. This confuses the cause and effect. This may make sense if some set of unique causes are responsible for the phenomenon of resilience. In this case, the metaphor of resilience comes to the fore. In the same paper as Southwick's definition, came the definition from Bonanno:

> "we define resilience very simply as a stable trajectory of healthy functioning after a highly adverse event" (Southwick et. al. 2014, p. 2).

Alternatively, in the same paper, Masten says:

> "resilience refers to the capacity of a dynamic system to adapt successfully to disturbances that threaten the viability, the function, or the development of that system" (Southwick et. al., 2014, p. 4).

Interestingly, the above definitions could apply to individuals or groups or even ecological systems (not necessary people). Left undefined are many terms that resilience definitions tend to hinge ("trajectory", "stable", "adverse event", "process", "capacity"). In yet another definition from the same paper, Panter-Brick says:

> "Resilience is a process to harness resources to sustain well-being. I like the word "process" because it implies that resilience is not just an attribute or even a capacity" (Southwick et. al., 2014, p. 5)[3]

The main point of going through this process of listing definitions is not to harp on problems with the definitions themselves (for now – but more in the next chapter), but to point out that several scholars in the same area are defining a term very differently (with some overlap) or with different terminology. This means, that in the first case, the phenomena demarcated by the APA with term *resilience* is different from the phenomena described by Bonano, Masten, and Panter-Brick while being put in contrast with each other all the same. In other words, these scholars in resilience could all be studying something completely different from each

[3] The idea that a process is not a property is not necessarily the only interpretation. Some explanations of mechanisms to appeal to properties, if one means by process, something like a mechanism (Craver & Tabery, 2019).

other, but none would be incorrect. How might this be resolved? I would like to pitch, in fact, that there is no resolution, because there is no real problem, only a Wittgenstein-esque linguistic puzzle. The linguistic puzzle of our own creation here is the use-referent or use-mention problem. The scholars are confusing the word for what the word demarcates. For example, in the phrase:

water has the letters “e” and “r” and is liquid

we are not separating the term *water* from what the term *water* refers to in the world (e.g., that which runs through streams). Panter-Brick in the aforementioned paper laments the lack of methods for measurement of resilience. However, the use-mention failings mean that we are confusing how to use the term *resilience* with what the term refers to. Arguing about what the term should refer to is an example of stipulative definition of which Hibberd (2019) notes is not productive for scientific definition. In other words, because of these problems, if we were to try to measure *resilience*, we might be in an eternal tug-of-war debating about how to use the term *resilience* while thinking we are investigating some phenomena to measure “it.”

To contrast with the definitional woes of NAEP, which seems to be about trying to answer some empirical questions about what makes someone read well and what relevant observations might be for measuring reading comprehension, resilience work seems to be tongue tied by linguistic conflations. This is not to say NAEP’s definition of reading comprehension is not to some extent stipulative (you can think of stipulative definitions as assigning meaning to a word ignoring other meanings or uses), but the work in resilience seems to be more about assigning meaning to the term. I hope that using two different research areas debating the definition of a primary property of interest can make more apparent how a seemingly similar efforts for the sake of measurement can be very different. In that sense, the questions, “what is reading ability?”, “what is resilience?” or “what is an atom?” are all very different questions.

Primarily, these two research areas can be used as examples for:

1. Why definition of a property of interest and conceptual analysis of that property is necessary for measurement, and the first required step in a measurement effort. Here, the term *property* will be used to denote that which we would like to measure – for instance, the height of a person or the person’s reading ability.
2. Productive and unproductive definitional work for the sake of measurement in the human sciences. What sorts of tasks are being accomplished with definitional work? The problems being solved in NAEP reading framework debates, resilience debates, and how they can be used to inform research hinge on definitions. How might we answer questions such as “what is reading comprehension?” Or “what is resilience?”
3. Both realms of research above have used latent variable modeling to measure. What can latent variable modeling provide and where is it detrimental?
4. Finally, what might be some criteria for good definitions? How do we define phenomena in such a way that these definitions are investigable and alterable?
5. Ultimately, I hope to use these examples to show that measurement and trust in measurement, is not merely about methods, checklists, or rule following (though, all may be relevant at some point), but a human process that involves sometimes frustrating back-and-forth conversations about that which is being measured and how we trust experiences of that which is being measured.

While this dissertation is not explicitly about solving problems in these two realms, I do aim to provide some tools and thinking that may help across many settings. I hope that having two concrete, real world running examples that are also areas of interest to this author may be helpful in clarifying the ways words may bewitch us, how this relates to psychometrics, and how we may use empirical means in some settings but not others.

1.3 Philosophy in science

Philosophy of language and philosophy and history of science will creep into this dissertation. Focusing on philosophy can feel like a frivolous mental occupation. After all there is data to analyze, studies to organize, and empirical papers to write as we all bow to the primacy of empirical work in education and psychology (a discussion of this primacy is given in Machado et al., 2000). Yet, all our work in the empirical social sciences require us to make sense of data. This sense making is rooted in a belief that we can take observations, imperfect as they may be, and make inferences beyond the setting of those data.[4] Claiming that we can learn about the latent causal forces of the world around us is, by its very nature, a philosophical framework that has not always existed nor is it presently universally accepted. If a researcher is using statistics, selecting among a suite of possible criteria for determining the fit and hence adequacy of a statistical model for some end, these fit statistics or criteria are based on some framework (and philosophy) of what sorts of features we would like our models to meet (for instance, adjustments for the complexity of a model is an orientation toward believing good theories or models are parsimonious; e.g. Schultz, 2018). The views of our present practice of statistics might be traced back to the likes of 16th or 17th century philosopher Francis Bacon (if we still want to stay in the last several centuries; but we can go

[4] Alternatively, if we reject that we can or should do that, this too, is a philosophical orientation.

back to Plato or his student Aristotle if we must; Mulaik, 1985 provides a nice historical account attempting to link certain forms of exploratory data analysis and Baconian traditions).

Doing philosophy is not just a realm for academic philosophers nor does it require immediately knowing how to interpret the tough jargon of philosophy. Philosophy is, in general, something we are already doing. However, just as a burgeoning statistician may not know what assumptions are required, a researcher making claims about measurement or theories of mind, this author included, may not always know the extent or strength of assumptions they're (I'm) making. Philosophy is often an attempt to make clear, even if not perfectly, why certain ideas or questions may sit at the tip of our tongues, especially in social science, but never quite feel clearly expressed or sensical. It is often an attempt to reveal contradiction or shape research questions into a coherent, answerable form, while laying bear what assumptions might be necessary or are implied by those research questions. In another sense, sometimes, we want to realize when we are asking the wrong questions or giving a "correct answer" to the wrong question.[5] The Statistician Bruno de Finetti, artfully described the interplay of science and humans:

> "Once the cold marble idol has fallen in pieces, the idol of perfect, eternal and universal science that we can only keep trying to know better, we see in its place, beside us, a living creature, the science which our thought freely creates" (de Finetti, 1989, p. 179).

[5] The status of the answer, "42," is well known.

I see this as an invitation to continually shape the practice of science in our respective fields, especially those fields as unsettled as the human sciences. In the section above, the idea that measurement is about human action, falls firmly in this category – our foundations make us blind to alternatives.

To channel de Finetti is to say that science is our own creation, and we are not passive participants. Some might argue that this is absurdly academic, that science must happen, and we need rigid rules for doing so to move quickly. This is what cutoff criteria for model fit or Bayes Factor cutoffs or the classic p-value of .05 provide us. R.A. Fischer, the very person who bestowed upon us p-values cutoffs of .05 would disagree with this charge, countering that this cutoff culture is in fact what is academic. He said that commitment to a strict cutoff is “absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he [sic] rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas" (Fisher, 1956, p. 42). A simple thought experiment in which different high stakes scenarios are involved, such as life or death, one might rely on more stringent cutoff values vs other scenarios – clearly, at this point, ritualistic practice of science can get in the way of doing science. I do acknowledge that it is hard to imagine a scenario that clearly directs us toward an appropriate cutoff value. Instead, we should view the cutoffs as less strong indications of anything than we usually do - for instance acknowledging something on one side of a p-value may not be qualitatively different from something just barely on the other side of the p-value.

The concepts presented below then, are meant to describe some useful ideas from philosophy in an inviting way (some things are perhaps dense, but I hope they are not impenetrable) and then I aim to integrate some of these ideas into empirical research practice.

In this realm, there are few things so nebulous as discovering or measuring psychological entities like “reading ability”, “math ability”, “depression”, and “resilience.” Some of this discomfort has been marginalized out, as discussed above, by implementing algorithmic rules for defining and measuring. Much of this dissertation will be devoted to the idea of conceptual analysis that asks what aspects of reality our words demarcate when we intend to measure something (“reading comprehension” refers to what in the world?). However, I cannot see this as a distinct enterprise from data collection – especially in education and psychology where modes of observation are generated in the form of test or survey questions. Where else can the relevance of these questions and the resultant observations come from but researcher thoughts, observations, and intuitions? The goal of conceptual and philosophical analysis is a form of transparency. Leaving something relatively undefined or closed to any form of modification in a scientific enterprise but making data and code freely available, is only a partial commitment to transparency. In a field like education or psychology, which do affect many people, it is too costly not to attend to the concepts that drive our research – it is, in fact, unethical.

1.4 Discussion and structure of the dissertation

This dissertation will use two running examples: one from the changing definition of reading ability for the NAEP reading report card, and one from an area of research in psychology called *resilience*. The purpose of this is to provide accounts of different ways that language may, in the words of Wittgenstein, bewitch us, and discuss how psychometric methods may or may not help, in some cases turning to the field of metrology for assistance, and finally, provide a brief example of a workflow that might integrate definitional or conceptual work with empirical work. For the most part, in this dissertation, uncertainty will be used colloquially as *not knowing*, but some initial formalizations will be provided drawing on work from the measurement sciences, specifically, metrology. This dissertation will focus on using basic ideas from the *philosophy of language,* metrology, and ontology to better unravel questions that commonly appear in research group meetings about measurement such as “what is resilience?” or “how should we measure a student’s reading ability?” These matters are tough because we are defining via introspection, which William James famously said is like “turning up the gas quickly enough to see the darkness” (James, 1890, p. 150).

While this dissertation will deal with educational settings primarily, its application is not meant only for education. It is meant, effectively, to be a toolkit for educational researchers and psychologists who struggle with questions of definition beyond operational definition and to hopefully corner some of the challenges with interpreting statistical results. Sometimes, though, the answer will not be satisfied with a hard and fast rule beyond “there will be no hard and fast rule for making decisions about interpreting results from research in human sciences.” Definition will be pitched as a partly linguistic, social, and empirical enterprise. Further, definition for the sake of measurement will be the framework of interest. Therefore, this dissertation has multiple, interrelated goals.

In Chapter 2, I will attempt to make the case for needing to do definitional work that is conceptual and not just empirical, that is at times informal and other times formal – or, really, transparent. In chapter 2, it will be argued that, in social science, partaking in definitional investigation is a scientific and ethical imperative. In other words, the nature of what is defined in the human sciences for the sake of measurement is not just a value-less matter. That which is defined for the sake of measuring may have real world consequences for large swaths of people – therefore the ethical and scientific are merged. Further, I’d like to argue that classic texts and guiding documents that have set forth the path for measurement in human sciences provide few if any tools for defining, so we likely need to make our own tools and guidelines. We need to know how definition of a property and the property itself fits into measurement as a process, so the structure of a measurement process from a psychometric and metrological perspective will be introduced. A brief discussion of how measurement happens and the role of homing in on the measurand, or what we hope to measure, will be used to introduce the field of metrology and its supporting documents.

The aims of chapter 3 will be to develop some useful frameworks and tools, largely drawing on the philosophy of language, for defining what is intended to be measured. This will include an understandable introduction to a particular realm of philosophy of language and connect this to measurement. The running examples of reading comprehension ability and resilience will be used throughout – applying this thinking to both realms. This will include a discussion of the use of the term "latent variable" which is common in psychometrics. But this requires some conceptual house cleaning about philosophy of science for words to refer to something and be relevant to science. Further, this will discuss how we might improve our beliefs about what is measured and what is not the aim of measurement, using language from a combination of coherentist and realist philosophies of measurement. However, I would also like to introduce a tension here between (1) certain forms of pragmatic realism which may admit certain things like psychological attributes as real despite being considered social constructions but also measurable and (2) the societal and power structures that lead to us considering them socially constructed but real. In other words, this can be a take on fairness in educational and psychological measurement that is specified typically via differential item function analyses. I do not propose a once-and-for-all solution to ethical issues to measurement in social science. Instead, I propose the idea of transparency as a possible first step in trying to be ethical about measurement.

Chapters 4 and 5 will discuss how we might start building consensus empirically and not just conceptually around definitions of properties. I shall use notions of differential item functioning or measurement noninvariance to explore and hypothesize about the causal features of the measurement process that are not intended to be measured. Hence, this would be an empirical means to help with definitional efforts. Language from metrology, especially around uncertainty and measurement error will be especially useful and informative for expanding linguistic frameworks in psychometrics. Explaining measurement noninvariance, (Zumbo, 2007; Zumbo et al., 2015) will instead be a goal for reasoning about what should and should not be considered in the definition of the property we'd like to measure as opposed to just an item or test altering tool. However, this demands a methodology. An example of the sorts of analyses and the reasoning behind it will be provided using a combination of latent variable mixture modeling and item response theory, while combining them with conceptual reasoning. Using theoretical and empirical work from a reading example, I hope to articulate, extending into chapter 5, an example of what iteration might look like to help develop definitions of properties to be measured.

Wrapping up, I aim to discuss and combine how empirical and conceptual work in psychometrics are not, or should not be, separate if measurement is truly an aim of psychometrics. That substantive psychological theory may be a part of psychometrics is not always well accepted in the psychometric community. Wijsen and Borsboom's (2021) interviews of past Psychometric Society presidents revealed a divide in perspectives about whether psychometrics is about statistics, psychology (here, lumping educational work, like testing, under the broad swath of psychology), or both. Wijsen and Borsboom (2021) noted that former Psychometric Society President Susan Embretson (notable for, among other psychometric work, focusing on psychometrics for the purpose of modeling learning and cognitive processes; e.g. Embretson, 1992; Embretson & Gorin, 2001) argued that "psychometricians can sometimes be too involved with technical details, whereas they should pay more attention to what they can contribute to psychological research" (Wijsen and Borsboom, 2021, p. 334). Wijsen and Borsboom directly quoted Embretson as saying that there are technically concerned psychometrician that "might be dealing with rather narrow statistical issues that are not really going to make a difference in the discipline" (Wijsen and Borsboom, 2021, p.334 – this is a direct quote of Embretson unlike the paraphrase in the sentence prior). Sijtsma and De Boeck also expressed views committing psychometrics as being *for* thinking about substantive issues for building theories. But this position is not for everyone. Paul Holland is quoted by Wijsen and Borsboom as arguing that psychometrics is really a branch of statistics, not so much psychology at this point in history. Wijsen and Borsboom (2021) note that one could argue that psychometrics "has lost its `psycho`-affiliation throughout the years and became a type of modeling that is relevant for a variety of research domains" (p. 335). Indeed, using statistics for research is tough as it becomes more apparent that off-the-shelf statistical models used in a variety of situations require forms of testing to make sure, for instance, that estimators are unbiased and efficient in unique settings, and undoubtedly, require some knowledge of probability and mathematics to know when those statistical models may lead us into trouble. This is firmly a statistical task. However, generating observations of people that we think are relevant for making inferences about people, thinking about the form or structure of the attributes we are hoping to make inferences about, are not firmly statistical but come from a range of research methods (anthropology and ethnography, informed participants in different settings such as teachers who have, often, unarticulated expertise, psychological and educational theories, etc.). This is not the realm of the statistician. Clearly, collaboration is needed, but, if one is to make a claim about having measured, to return to the intro, we need to have strong positions about having measured *something.* I do not see how measurement can then purely be in the realm of the statistician. There is little to no mapping between theories of point estimation and psychological theories aside from those smuggled in by early statisticians, who, for many parts, used a theory of errors to introduce eugenicist thinking (e.g. Mulaik, 1987). Thus, it is the position of this dissertation that if psychometrics is not about psychology and is only about statistics, psychometrics cannot be about measurement. It is fine if psychometricians want to take a statistical turn – but then measurement is a separate effort. Nonetheless, I hope the final portion of this dissertation will connect, to some effect, the way that empirical statistical work may guide definitional efforts.

References

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425. https://doi.org/10.1007/S11336-006-1447-6

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Connor, K. M., & Davidson, J. R. T. (2003). Development of a new resilience scale: The Connor-Davidson Resilience Scale (CD-RISC). *Depression and Anxiety*, *18*(2), 76–82. https://doi.org/10.1002/da.10113

Craver, C., & Tabery, J. (2019). Mechanisms in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2019). Metaphysics Research Lab, Stanford University.

de Finetti, B. (1989). Probabilism: A critical essay on the theory of probability and on the value of science. *Erkenntnis,* 169–223.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155.

Embretson, S. E. (1992). Measuring and validating cognitive modifiability as an ability: A study in the spatial domain. *Journal of Educational Measurement*, *29*(1), 25–50.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343–368.

Fechner, G. T. (1987). Outline of a new principle of mathematical psychology (1851)*. In *Psychol Res* (Vol. 49).

Finn, C. E. Jr. (2021, May 2021). The culture wars come for the Nation's Report Card | The Thomas B. Fordham Institute. *Fordham Institute Blog*. https://fordhaminstitute.org/national/commentary/culture-wars-come-nations-report-card

Fisher, R. A. (1956). *Statistical methods and scientific inference.*

Flake, J. K. (2021). Strengthening the foundation of educational psychology by integrating construct validation into open science reform. *Educational Psychologist, 56*(2). https://doi.org/10.1080/00461520.2021.1898962

Gehlbach, H., & Hough, H. J. (2018). Measuring Social Emotional Learning through Student Surveys in the CORE Districts: A Pragmatic Approach to Validity and Reliability. *Policy Analysis for California Education, PACE*.

Giordani, A., & Mari, L. (2012). Measurement, models, and uncertainty. *IEEE Transactions on Instrumentation and Measurement*, *61*(8), 2144–2152.

Haig, B. D., & Borsboom, D. (2008). *On the Conceptual Foundations of Psychological Measurement*. https://doi.org/10.1080/15366360802035471

Hibberd, F. J. (2019). What is scientific definition? *Journal of Mind and Behavior, 40*(1). https://psycnet.apa.org/record/2019-31815-002

James, W. (1890). The principles of psychology, Vol I. In *The principles of psychology, Vol I*. Henry Holt and Co. https://doi.org/10.1037/10538-000

Joint Committee for Guides in Metrology (JCGM). (2012). *International Vocabulary of Metrology—Basic and general concepts and associated terms (VIM)* (3rd ed.). JCGM. (2008 version with minor corrections). www.bipm.org/en/publications/guides/vim.html

Joint Committee for Guides in Metrology (JCGM), Jcgm, J. C. F. G. I. M., Bipm, Iec, Ifcc, Ilac, Iso, Iupac, Iupap, Oiml, Jcgm, J. C. F. G. I. M., Joint Committee for Guides in Metrology, Jcgm, J. C. F. G. I. M., Drive, B., Issue, F.--, Drive, B., Williams, T., Kelley, C., Campbell, J., … Parkway, W. M. (2008). Evaluation of measurement data

— An introduction to the “Guide to the expression of uncertainty in measurement” and related documents. *International Organization for Standardization Geneva ISBN, 3*(October).

Knight, G. P., Roosa, M. W., & Umaña-Taylor, A. J. (2009). *Studying ethnic minority and economically disadvantaged populations: Methodological challenges and best practices.* American Psychological Association.

Kolen, M. J., & Brennan, R. L. (2014). Linking. *Test Equating, Scaling, and Linking*, 487–536. https://doi.org/10.1007/978-1-4939-0317-7_10

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/SCIENCE.AAL3618/ASSET/108C1BB1-018D-4B37-B6C6-2520A41FA197/ASSETS/GRAPHIC/355_584_F1.JPEG

Machado, A., Lourenço, O., & Silva, F. J. (2000). Facts, concepts, and theories: The shape of psychology’s epistemic triangle. *Behavior and Philosophy,* 1–40.

Mari, L. (2013). A quest for the definition of measurement. *Measurement: Journal of the International Measurement Confederation*, *46*(8), 2889–2895. https://doi.org/10.1016/j.measurement.2013.04.039

Mari, L., Carbone, P., & Petri, D. (2012a). Measurement Fundamentals: A Pragmatic View. *IEEE Transactions on Instrumentation and Measurement*, *61*(8), 2107–2115. https://doi.org/10.1109/TIM.2012.2193693

Mari, L., Carbone, P., & Petri, D. (2012b). Measurement fundamentals: A pragmatic view. *IEEE Transactions on Instrumentation and Measurement*, *61*(8), 2107–2115. https://doi.org/10.1109/TIM.2012.2193693

Mari, L., Maul, A., Torres Irribarra, D., & Wilson, M. (2017). Quantities, Quantification, and the Necessary and Sufficient Conditions for Measurement. *Measurement: Journal of the International Measurement Confederation*, *100*, 115–121. https://doi.org/10.1016/j.measurement.2016.12.050

Mari, L., Maul, A., & Wilson, M. (2019). Can there be one meaning of “measurement” across the sciences? *Journal of Physics: Conference Series*, *1379*, 12022. https://doi.org/10.1088/1742-6596/1379/1/012022

Mari, L., Wilson, M., & Maul, A. (2021). *Measurement across the sciences: developing a shared concept system for measurement*. Springer.

Marsh, J. A., McKibben, S., Hough, H. J., Allbright, T. N., Matewos, A. M., & Siqueira, C. (2018). Enacting Social-Emotional Learning: Practices and Supports Employed in CORE Districts and Schools. *Policy Analysis for California Education, PACE*.

Maxwell, G. (2009). The ontological status of theoretical entities. *Philosophy of Science: An Historical Anthology*, 451–458.

Meehl, P. E. (1992). Factors and Taxa, Traits and Types, Differences of Degree and Differences in Kind. In *Journal of Personality* (Vol. 60). https://meehl.dl.umn.edu/sites/g/files/pua1696/f/150factorsandtaxa.pdf

Messick, S. (1995). *Validity of Psychological Assessment*. https://search-proquest-com.proxy.library.ucsb.edu:9443/docview/614327710/fulltextPDF/C82B49CF016A4D1FPQ/1?accountid=14522

Michell, J. (1999). Measurement in Psychology: Critical History of a Methodological Concept. *Measurement in Psychology a Critical History of a Methodological Concept*. https://doi.org/10.1017/CBO9780511490040

Michell, J. (2003). The Quantitative Imperative: Positivism, Naive Realism and the Place of Qualitative Methods in Psychology. *Theory & Psychology*, *13*(1), 5–31. https://doi.org/10.1177/0959354303013001758

Mulaik, S. A. (1985). Exploratory statistics and empiricism. *Philosophy of Science*, *52*(3), 410–430.

NAGB. (2021, March 4). *Reading Comprehension in Large-Scale Assessment: A Symposium*. https://www.nagb.gov/naep-subject-areas/reading/results-archive/reading-comprehension-symposium.html

National Assessment Governing Board. (2021). Reading Framework for the 2026 National Assessment of Educational Progress. In *NAEP*.

Nitko, A. J. (2016). Distinguishing the Many Varieties of Criterion-referenced Tests: *Http://Dx.Doi.Org.Proxy.Library.Ucsb.Edu:2048/10.3102/00346543050003461*, *50*(3), 461–485. https://doi.org/10.3102/00346543050003461

Schultz, M. (2018). The Problem of Underdetermination in Model Selection: *Https://Doi-Org.Proxy.Library.Ucsb.Edu:9443/10.1177/0081175018786762*, *48*(1), 52–87. https://doi.org/10.1177/0081175018786762

Schwartz, S. (2021, August 13). “Nation’s Report Card” Has a New Reading Framework, After a Drawn-Out Battle Over Equity. *Education Week*. https://www.edweek.org/teaching-learning/nations-report-card-has-a-new-reading-framework-after-a-drawn-out-battle-over-equity/2021/08

Scribner, S. (1984). Literacy in three metaphors. *American Journal of Education*, *93*(1), 6–21.

Southwick, S. M., Bonanno, G. A., Masten, A. S., Panter-Brick, C., & Yehuda, R. (2014a). Resilience definitions, theory, and challenges: interdisciplinary perspectives. *European Journal of Psychotraumatology*, *5*. https://doi.org/10.3402/ejpt.v5.25338

Southwick, S. M., Bonanno, G. A., Masten, A. S., Panter-Brick, C., & Yehuda, R. (2014b). Resilience definitions, theory, and challenges: interdisciplinary perspectives. *European Journal of Psychotraumatology*, *5*(1). https://doi.org/10.3402/ejpt.v5.25338

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*(2684).

Tal, E. (2016). Making Time: A Study in the Epistemology of Measurement. *Https://Doi-Org.Proxy.Library.Ucsb.Edu:9443/10.1093/Bjps/Axu037*, *67*(1), 297–335. https://doi.org/10.1093/BJPS/AXU037

Tal, E. (2019). Individuating quantities. *Philosophical Studies 2019 176:4*, *176*(4), 853–878. https://doi.org/10.1007/S11098-018-1216-2

U.S. Department of Education. (2010). *An Introduction to NAEP: National Assessment of Educational Progress*.

Wagnild, G. M., & Young, H. M. (1993). Development and psychometric evaluation of the Resilience Scale. *Journal of Nursing Measurement*, *1*(2), 165–178. http://www.ncbi.nlm.nih.gov/pubmed/7850498

Wijsen, L. D., & Borsboom, D. (2021). Perspectives on Psychometrics Interviews with 20 Past Psychometric Society Presidents. *Psychometrika*, *86*(1), 327. https://doi.org/10.1007/S11336-021-09752-7

Windle, G., Bennett, K. M., & Noyes, J. (2011). A methodological review of resilience measurement scales. *Health and Quality of Life Outcomes*, *9*(1), 8. https://doi.org/10.1186/1477-7525-9-8

Zumbo, B. D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). *A Methodology for Zumbo's Third Generation DIF Analyses and the Ecology of Item Responding*. https://doi.org/10.1080/15434303.2014.972559

Chapter 2
The Ethical Imperative

To compare measurement results across contexts, we need to articulate and agree on what is being measured with a given measuring instrument. For instance, we might need specific tools to measure length in one setting but a different set of tools to measure length in another. Chang (2004) describes how new tooling was needed for measuring very hot or cold temperatures. Measuring distance in different settings with different tools might be required in transportation, for instance, where there are varying and nonconstant dimensions a road may traverse. In this dissertation, it is taken as uncontroversial that one particular purpose of measurement is gaining knowledge about something in a particular way and that measurement results are meaningful when they can be interpreted across context (e.g., understood intersubjectively as in Mari et al., 2012a). In other words, measurement is useful to science when measurement results can be replicated across contexts *and* it is known that the source of replication is that the same thing is being measured each time. A logical result of this is that when there are differences in measurement results across contexts, we can advance by questioning the quality of measurement in each setting or questioning, for instance, what might be the causal mechanism that is changing results across contexts (Tal, 2019). Perhaps the thing we are hoping to measure is a composite of many attributes, some of which we may not be interested in. Alternatively, perhaps something aside from what we want to measure is changing our measurement results.

Focusing on understanding what one is measuring and paying attention to the definition of the property under study is important for the sake of *generalization*. It is important to know what is being measured using different (or similar) measurement operations across many contexts for the sake of comparability. I shall consider the properties under measurement to be *universals* – for now, we can define universals as non-arbitrary characteristics of attributes that many different objects may share. Multiple objects may share the property of height (for more thorough discussions of universals and properties, see Armstrong, 1980; Orilia & Paolini Paoletti, 2022a). As was noted above, it is maintained throughout this dissertation that a property of an object is what is measured, following Mari et al., (2021). For instance, in the height of a person – height is a property. Height in general, is a universal property meaning the same property height can be realized, albeit with the same or different values, across objects or persons. Two different people can have the same height value. In this sense, because height is a universal (and general property; Mari et al., 2021, chapter 5 and chapter 6), one can measure height in different settings. The property height in this case, will be generalized. This is perhaps easier to conceptualize in non-social or human settings.

According to some like Armstrong (1980), universals exist mind independently. However, this proves problematic for accepting universals in psychological measurement since psychology is the study of the mind, and many of the ideas for which we study are derived or requiring of some introspection of how we individually experience these ideas. One possible solution to this comes from the likes of Searle (1995) and Maul (2013). If measurement in education and psychology is possible at all and about properties, can those properties be considered universals? Searle sees the mind as part of the natural world and hence implies that it is something that can be studied. Yet, the mind is not necessarily spatially or temporally located – though, the relation between the mind and brain make it so. However, an appeal to intersubjectivity of conceptualizations of ideas, properties, or things (that money has value) made by Searle (1995) and Hacking (1999), place existence of a psychological attribute (for instance, reading comprehension) not in a single person's mind, but as socially constructed, intersubjectively understood, and having an influence on people's lives. In this sense, there is a sort of pragmatic realist position that psychological attributes can be real because it would be odd to not treat something like physical objects constructed by humans as real (e.g. – the University is only a University as opposed to a particular collection of buildings because of the human activities classified as the things done at a university). This admits the possibility that psychological attributes could in principle be real or considered universals. However, what gets treated as a universal then becomes a delicate matter of power wielding. The psychometricians control the narrative of what is measured for instance and what is measurable in the human sciences (Borsboom & Wijsen, 2017; Wijsen et al., 2022). So, while this form of realism admits one possibility, it introduces other ethical problems.

Containing any effort to define phenomena in the human sciences to an opaque (or black) box that is logically and empirically inflexible may be a form of epistemic injustice, and, just generally, poor scientific practice. In other words, it is maintained here and throughout this dissertation, that what is often considered peripheral to measurement in the human sciences, namely, sociocultural and linguistic concerns, are actually core to identifying and explaining the phenomena we observe such as reading activities or persons' propensities to think through, adapt to, and act on challenges (e.g. *psychological resilience)*. Two subsets of epistemic injustice, "hermeneutical injustice" and "testimonial injustice" (Fricker, 2007) are especially important. Hermeneutical[6] injustice precedes testimonial

[6] Here, we might say hermeneutics are about human interpretation of the world and understanding of the world. For instance, how we interpret the world changes our understanding. One could argue that

injustice and is the case when someone is "at an unfair disadvantage when it comes to making sense of their social experiences" (Fricker, 2007, p.1). The idea here may be that there is no set of concepts or words to represent a person's experience. Testimonial injustice occurs when "a hearer gives deflated credibility to a hearer's word" (Fricker, 2007, p.1). Fricker places these injustices in the hands of the state, or, at least, in the hands of those in power. This terminology is useful for establishing a framework for fairness in educational and psychological measurement.

An example of where some might be concerned with hermeneutic and testimonial injustice concerns are the legitimate definitional debates surrounding the National Assessment of Educational Progress (NAEP) reading assessment. As mentioned in chapter 1, the NAEP Framework names "expanding the construct of reading" (National Assessment Governing Board (NAGB), 2021, p.3) as a primary goal for the new NAEP reading framework. It acknowledges that "research evidence has highlighted that … reading comprehension is a meaning-making activity that involves socially and culturally specific characteristics and practices" (NAGB, 2021, p. 4). This is an expanded *definition* of reading comprehension that many scholars on NAEP's visioning panel claim is a more fair but also a more accurate view of reading. It is written in the 2026 NAEP Reading Framework that a new framework is desired to "update the framework in a manner that would enhance the assessment's validity and fairness while minimizing bias[7]" (NAGB, p. 4). It seems intrinsic

much of social or the human sciences is hermeneutical since it is about making sense of the world around us (George, 2021)

[7] It is not exactly clear how the visioning panel relates fairness to bias. If one were to guess – both terms are used in a psychometric sense. Namely, that fairness is a matter of minimizing differential item functioning such that item response probabilities are equal no matter a person's group membership while bias, or test bias, occurs at the test level when test scores are over or underestimated for certain groups (see, for instance, Zumbo, (2007)Zumbo, 2007)Zumbo, 2007).

that NAEP's visioning panel, by trying to incorporate readers' different senses of the world may be concerned with hermeneutical injustice issues. If we take a reading scenario as a particular instance with fuzzy boundaries then the way students read is by integrating their sense of the external-to-the-reading-passage world and utilizing concepts from their own world to bring into the reading world. Alternatively, if one were to imagine a NAEP testing situation as a closed off box, the reality of that testing situation will be devoid of "conceptual resources" (Dunne, 2020, p. 4) for some students. Fricker (2007) uses a more poignant example – that of sexual harassment which only seemed to be a concept realized in the 1960s and 1970s. Up to that point, there would be few ways or reasons to even consider certain actions *as* sexual harassment (at least not clearly) and hence problematic. Fricker says women in this case were hermeneutically marginalized – when a group (or groups) of people are offered little or unequal (in a negative sense) interpretive power of the world around them. Fricker (2007) thus expands the definition of hermeneutical injustice:

> "the injustice of having some significant area of one's social experience obscured from collective understanding owing to a structural identity prejudice in the collective hermeneutical resource" (p. 156).

Epistemic injustices then are clearly relevant to the matter of defining that which we want to measure and selecting what to measure in education and psychology, especially if we admit socially constructed entities (and what's not, at the end of the day?). Definition is a concern about delimiting ontological commitments - ontological commitments are field or even person specific commitments or beliefs about ways things exist. This clearly turns into a scientific or epistemic concern and not external to the research-based enterprise.

Admitting and allowing more groups of people to make sense of reality are indeed concerns of science when those experiences are the subject of study. As mentioned above, Hacking (1999) and Searle (1995) note that just because something is socially constructed (e.g. – the concept *woman refugee* from Hacking) does not mean these things should not be considered real since the labels act on and are absorbed by bearers of those labels, changing how the bearer of the label acts. Hacking (1999) claims that something being called *socially constructed* is often an *unmasking*. Unmasking is pointing out that the way the world is does not have to be so but, at the same time, this is indeed how the world is. In this sense, defining something like reading comprehension ability through psychometric avenues gives certain groups – likely academic institutions or testing organizations – more authority.

In Fricker (2007) as well as (Teo, 2018, Ch. 4), the notions of epistemic injustice (Teo might phrase some version of this, *epistemic violence*) are matters of power. Fricker argues that power is a coordination among people, even if tacit. For instance, universities grade students in a particular way and have power because many employers rely on grades and grading even if indirectly, to select employees, thus increasing the power of the instructor in a class (Fricker, 2007, p. 12, drawing from Wartenberg, 1992). Meanwhile, the words we use to define are governed by a certain conceptual system. It is hoped that throughout this dissertation, I can make clear that language is important for demarcating what we aim to study but that defining and the entities that are thought to exist (e.g. reading comprehension ability, resilience), especially in the human sciences, need to be left open for probing**.**

2.1 Ontological Commitments and Ethics

In the introduction chapter, it was argued that, in essence, settling on a satisfying definition of a property for measurement is a matter of agreement (or, in this case, hermeneutics). However, this introduces the tension between the socially constructed as real and the necessity of identifying and agreeing on the nature of what is measured. While in all sciences, power structures dictate what is studied, in the human sciences, the agreed upon properties that we study are properties and attributes of people (as well as the relations among these things). Being of people means that we may bring into existence only certain things but those things, as noted above, can be absorbed by bearers of the property labels. This necessitates the need for transparent definition that is revisable and believed to be fallible, since these socially constructed entities, like psychological properties, and the way they may manifest, requires the input of the people they affect to avoid hermeneutical injustices. Alternatively, what isn't studied, can have real effects on people as well. Therefore, ontological commitments, even from a modeling perspective need to be laid bare.

Thus, introducing uncertainty about a definition, which in this case means specifying the ways we do not know what *something* is, or just leaving that definition to be ever revisable is a matter of fairness and scientific effort. Considering that measurement in psychology and education have roots in eugenics, it seems that caution is warranted about admitting everything as measurable (with a lineage[8] of eugenicists, for instance, heading the Psychometric Society; see Wijsen & Borsboom, 2021). Some have noted that this eugenic history has directed psychometric thinking. Psychometrics "was considered a tool for setting up the "ideal" society, in which intelligence measurement would play an incremental role in

[8] Pun intended

setting up a meritocratic hierarchy, eliminating crime, and deciding who was encouraged to procreate and who was discouraged from doing so" (Wijsen et al., 2022). It is not a logical leap to see the closeness of this reasoning in college admissions, though, college admissions may have different goals – considering the effects of these goals would seem to measure how far we have come from those eugenic ideals. Psychometrics as a technical discipline, as will be discussed below, has had little to say, until relatively recently, about what is measured or what parameters in models refer to. But, indeed, tools like factor analysis and item response theory are used as if they refer to psychological entities.

Operationalism, in some settings, defines what might be called "traits" (and the eugenicists called traits – see, Noorgard, 2008). This equating of traits to factors in factor analytic models, however, means that there is no objection to the empirical or logical contents of what is measured – the thing that the measurement process defines in an operationalist setting. In an anything goes mentality, those who wield measurement tools (testing companies, psychological assessment owners) get to decide what matters for measurement. There is no room for transparency. When decisions are being made about people, the strength of evidence should be accessible such that an informed public, especially those that are influenced by these tools and decisions can ask, "what is the evidence and how strong is it? Is this something that should be measured?" For operationalism or formative measurement, little evidence is needed at all to justify measurement claims because measurement claims make no effort to stake out constituents of reality, there is nothing to defend about the adequacy of measurement. For instance, whether a "measuring" instrument appropriately measures something cannot be debated on coherence grounds because there is no *something* to stake out – it is coherent because it is defined by measurement operations. In the following sections and chapters, the focus is on the *definition* of what is to be measured as well as empirical methodology for doing so. More explicitly, the focus is providing a more explicit framework for examining and using everyday language for the purpose of defining as a starting place for measuring a given psychological attribute – the *measurand*.

Examples of this ethical imperative to devote time to definition are measurement claims in NAEP – the National Assessment of Educational Progress - about subgroups. NAEP's use in the United States for subjects such as reading are communicated via national reports. Presumably, and as discussed briefly in chapter 3, NAEP's history is related to a post WWII political desire for the United States to remain competitive on the international stage – so if students can be monitored, presumably, interventions could be used to improve student academic performances (a definition of what is measured is often lacking). NAEP reading, for instance, provides a set of things students should be able to do while reading. The NAEP reading framework (for now ignoring if this is indeed a good characterization), begins by describing what good readers do and cites, for instance, a RAND study definition of reading comprehension, a PIRLS definition of reading literacy, and a PISA description of what 15-year-olds can do via the term *reading literacy* as well. However, prior to the 2019 framework, NAEP describes reading skills. It is noted, that no single measurand is identified and leaves little room for objection, since some of these definitions over-admit or under-admit (depending on the perspective) what should be considered part of "reading ability" (NAEP Reading Framework, 2019). The 2025 (now 2026) reading framework is attempting to expand what constitutes reading or necessary reading skills.

If we are to analyze NAEP reading debates, there are some matters that are hard to interpret about the definition of reading from both sides of the debate outlined in the introduction. On the side of the reading debates that is cognitivist and sees reading ability measurement as focusing on the cognitive processes required of reading, they ask that certain words or concepts be potentially predefined for students to scaffold item responses. The implication is that this pre-knowledge is not part of the phenomenon of reading that this side of the NAEP debate thinks should be studied. In other words, sociocognitive perspectives on reading tacitly argue that the students' worlds and background knowledge may lead to changes in the outcome of the reading response but is not the core property of interest for measurement in reporting NAEP scores. Meanwhile, the other side of the debate admits background knowledge as part of the phenomenon of reading. Clearly, this is a value-based debate that might benefit from considering the nature of a psychological phenomenon as an entity. From an ethical perspective, the basic notion is that there is an apparent achievement gap between, as typically classified, white students' NAEP test scores and Black and Latino students' test scores, leading to claims about subgroup "abilities" and reasons for this discrepancy (e.g. Ladson-Billings, 2006). In principle, this could have implications when attributing these gaps to say, background knowledge differences or attributable to something related to cognitive processing at the level of the word or even lexeme or something more macro (e.g. Kintsch & van Dijk, 1978; not to misrepresent Kintsch and van Dijk who referred to micro structures of semantics as, at times, a higher level that individual words and morphemes to make sense of a text, though, the reading process would include the integration of decoding – making sense of individual words). In one conception, the target property is not inclusive of background knowledge, though background knowledge is an important causal feature in the measurement of reading comprehension that need be adjusted for. This casual structure is recognized in Figure 2.1a.

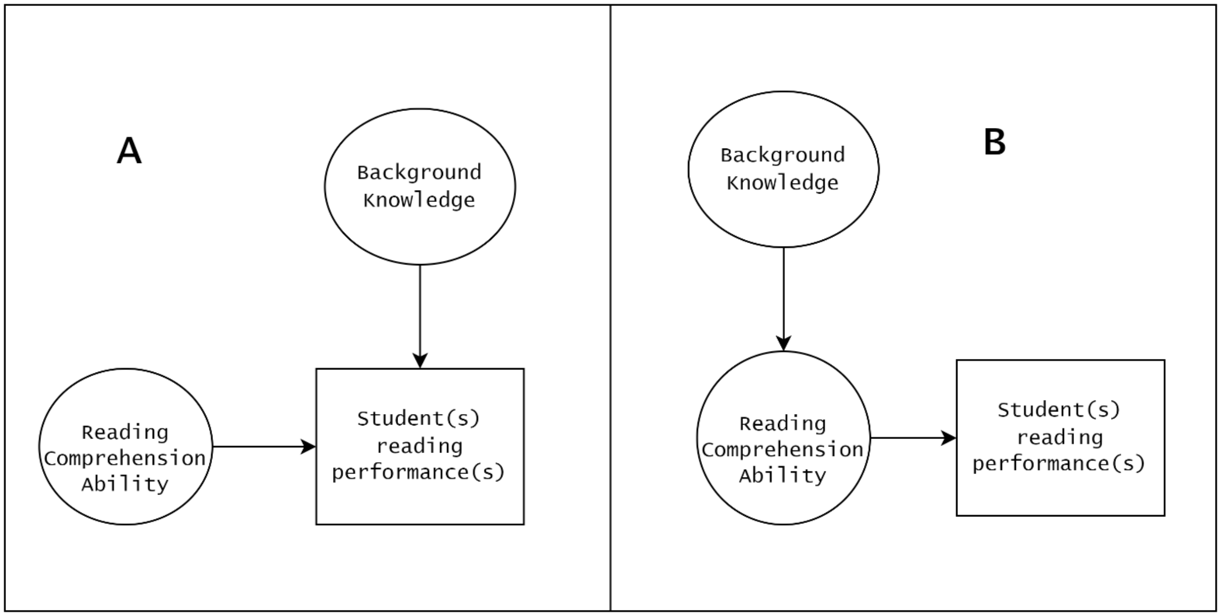

A
Background Knowledge
Reading Comprehension Ability
Student(s) reading performance(s)
B
Background Knowledge
Reading Comprehension Ability
Student(s) reading performance(s)

Figure 2.1a and Figure 2.1b. *In figure 2.1a, on the left - the property of interested is some form of reading comprehension ability that involves cognitive processes integrating information from the text that students have access to. Background knowledge also has an external effect on the reading outcome but is not the target of interest. 2.1b provides another plausible perspective where background knowledge is external but effects reading comprehension ability directly, as opposed to just the item response. Both seem consistent with the cognitivist side of the NAEP reading debates but would lead to very different inferences, items, or models and may be more or less compatible with the other side of the reading debate who want to integrate background knowledge in NAEP reading definitions.*

The vague nature of the definition of reading plus the general problem that many different theories can account for the same data patterns, even in the best settings, can lead to outsized, unfair, and racist claims about these subgroups without definitional qualification (to what extent is this conception going to influence how subgroups are treated?). Teo (2018, p.9) provides the example that psychology used to have a technical term “moron” and this term worked its way into U.S. policy to exclude certain individuals from jobs.

In education discourse, language and concept analysis is not new. For instance, Israel Scheffler in his book, (1963) *The language of education*, attempted to provide a case study in concept analysis focusing on teaching. Like Hibberd, Scheffler saw definitional work for the purpose of science as necessary. Invoking something between Wittgenstein and perhaps predicting Latour, Scheffler saw that “scientific definitions, in particular, are continuous with contemporaneous statements in their environing networks, and cannot well be evaluated in abstraction from these networks” (Scheffler, 1963, p. 12). He argues that definitions in science have a certain special meaning to their own scientific communities but can take on dangerous lives of their own when used to communicate to external audiences and that paying attention to the *type* of definition is important for understanding and clarifying ideas. He argues that sorting out what the aims of defining a term are and its lineage, will help also clarify the conceptual analysis attached to that word. To sum up, when questions such as “what is resilience?” are asked by researchers, this is a definitional task. But the answer to that question will depend not just on the way the world is (or we think it is), but what a given term means in a given community. As science changes, so do terms and their meaning. Conceptual analysis for the sake of science is not meant to arrive at a decision that is final – it is not believed possible anyway – but instead to help us get on the same page. Sometimes, we will think we are on the same page when we’re not because of using the same terminology in different wats. Conceptual analysis and philosophy of language can help sort this out. Therefore, part of this dissertation will be devoted to ontic commitments and the other will be devoted to epistemic concerns such as statistical methods for interrogating our ontic commitments.

2.2 Definitions of Properties in Classic Psychometric Texts

While Hibberd's (2019) writing provides some useful insights into what sorts of conditions are satisfied by good definitions – which I shall describe later - what might be desired still are some tools to use for definitional work for the explicit purpose of measurement. The lack of emphasis on definition has real world consequences. In the realm of educational assessment and certain realms of psychology, it is often difficult to pin down what a researcher or psychometrician or testing agency intends to measure in a way that is useful, say, for a teacher or curriculum developer. Even if one were to take a pragmatic perspective that fruitful science can be defined in terms of how new ideas or tools can enhance our interaction with the world – vague definition would not suffice. A vague definition would not direct actions clearly enough to give that definition any value. For instance, in the 2017 technical document for the SAT, there is no explicit discussion of what the SAT is measuring beyond "the SAT suite of products provide better information about students' strengths and weaknesses relating to the knowledge, skills, and understandings that are essential to college and career readiness and future success" (*SAT Suite of Assessments Technical* Manual, p. 9). The document uses phrases such as "apply their reading, writing, language and math skill to answer questions in science, history, and social studies" (p. 8) in describing what is required of students in the newly modified areas of "Analysis in Science and History/Social Studies" but does not make it clear if these skills are what are being measured, what the structure of these skills might look like (e.g. their cognitive content or their very natures, for instance, whether they should be modeled quantitatively at all or investigated via other means) or whether these are being measured distinctly or treated as some large composite. Taking a pragmatic perspective, it is not clear what one would *do* with these concepts, either, since there is no mapping from test content to particular actions for students. If one were to argue that the SAT is not claiming to measure anything but is merely generating data for useful prediction, this would be odd since the term *measure* (or measurement) is used at least 130 times in the document without an explicit definition of what is being measured beyond claiming to "assess the skills, knowledge, and understandings that matter for college and career readiness" (p. 108).

Unfortunately, as noted by Hibberd and Borsboom, even foundational psychometric texts seem unconcerned with the aspect of measurement devoted to property identification. Ignoring for the time being, their use of the term, "theoretical construct," the nearly biblical Lord & Novick (1968), say that an "observable variable is a measure of a theoretical construct if its expected value is presumed to increase monotonically with the construct" (Lord & Novick, p. 20). The inherent distinction here is that there are "variables" that are observable and some that are not, but there is no justification for this distinction. Instead, Lord & Novick (1968) are speaking of empty sets, two classes of variables, or entities, that are by their nature, observable or not. They have no particular property of interest in mind but speak of these "theoretical constructs" as if they are a special, unified, set. They note that the problem is "the concepts of theoretical interest [in social science] tend to lack empirical meaning…the more "theoretical" constructs are often not far removed from simple common sense" (Lord & Novick, 1968, p. 15 citing Torgeson, 1958). It is worth noting that their concept of observable and unobservable entities are non-technical even if invoking a statistical or mathematical conceptualization because they rely on common sense understanding of these terms.

Lord & Novick do not rely on a definition of what is being measured and instead turn toward mathematical formalization. They define a true score of a "theoretical construct" as an expectation. Given theoretical infinite replications, the true score is defined as the expected value, or mean score, of a person on a test (E[x] = true score – or $\int xf(x)$dx where x is the score and f(x) is some probability density (or mass) function and the integral can be replaced by a summation in practice). This move requires no commitment to a definition of a property because a true score is not what they call a *platonic true score* or even "construct score." A true score is the expectation of whatever score an instrument will assign to a person as opposed to what the person's actual value on a construct of interest is. They clarify: "a person's true score will depend on the various kinds of conditions under which the measurements are taken" (Lord and Novick, 1968). This would make sense if they were arguing for something akin to: *in different scenarios, a person changes, and the value of some mental state at that moment is different than it was in a different moment. For instance, if you turn up the heat in a room, a person might feel warmer than they did in the room an hour prior*. Or, they could argue: *The value a measuring instrument provides is different in different settings as things from different settings interfere with an ideal realization of a value for the property of interest.* However, Lord & Novick do not seem to mean this. Instead, they seem to mean that a person's true score is definable by condition – a new true score for each test or setting. This is a form of operationalism – the true score is defined by the way in which something is measured. Lord & Novick, then, are not the ones to turn to for definitions of properties to be measured and their justification for assuming one has measured something does not seem to hold water. Yet, the tenets of classical test theory (reliability, true scores, test or assessment level standard errors of measurement) remain as dominant methodology in psychology and education.

However, simple dismissal of foundational writings does not garner a useful approach for defining educational or psychological entities. One point of agreement among scholars seems to be that a challenge facing educational and psychological measurement and the human sciences generally is that they are full of what Maraun (1998) calls “common-or-garden concepts” which are “concepts with a common employment in everyday life” that are taught and learned by the *“person on the street”* (Maraun, 1998, p. 453). Maraun contrasts this with technical concepts which are defined by “specialized or expert” communities and these technical concepts have a “narrow, technical field of application” (Maraun, 1998, p. 453), much as Scheffler argued. This identification of the plague of common or garden concepts which have vague and multiple meanings rooted in metaphorical foundations of common language (Lakoff & Johnson, 1980) is not unique to Maraun. Lord & Novick present their theory of errors along with classical test theory and IRT models as a solution to the vague semantic components of psychological and educational theoretical work– they desired to develop a formal syntax where syntax would effectively be a statistical model. More recently, a push to formalize psychological theories (for which I am grouping notions of educational abilities such as “math ability”, “reading comprehension ability”, or similar) has manifested in a suite of papers that thoughtfully promote the use of more careful language rooted in formalization via mathematical, statistical, and computational modeling (for instance, see Borsboom et al., 2021; Fried, 2021; van Rooij & Blokpoel, 2020). To some extent, though, there is needed work in the realm of understanding how to talk about what they are modeling – the language will dictate how we model or predict from concepts.

2.3 Where else shan't we look for how to define?

2.3.1 The AERA, APA, NCME Standards and its validity theory:

Though a full discussion of validity is beyond the scope of this dissertation, *test validity* is such a prominent term in educational assessment that it would be hard not to address it. The opening of the first chapter of the 2014 *Standards for Educational and Psychological Testing* (hence forth, the *Standards)*, which is supposed to be abided by testing organizations, research institutions, or anybody who is crafting an academic or psychological test of some sort, states*:*

> "Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests. Validity, is therefore, the most fundamental consideration in developing tests." (American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA, NCME), 2014, p. 11)

The oddness of this definition is caused by pitching it as if it is a true consensus – using the term "refers" to convey an authoritative air about what the term validity means, as opposed to noting that it is indeed the authors' proposed definition (for instance, the authors could have written – "*we will use the term validity to refer to…*"). While some may argue that this is a consensus definition, others have argued that, if it is a consensus, it may be a weak one. One problem is that the term validity as used by test makers and designers or other measurement professionals is quite different from the term "valid" as used by policy makers, the general public, or even philosophers and mathematicians. For instance, a valid argument is very different from a valid test but some, like Kane, do connect validation and testing to argumentation. Baker (2013) notes that tests and assessments have many different uses ranging from professional certifications to identifying if students meet certain educational standards – effectively checklists - while other tests may operate more like measurements that give students scores that should, in principle, be interpreted. According to the *Standards,* validity would refer to any of the above, even if a test is just to show that a student can carry out certain skills (did a test taker perform this task?) which may require no property or construct to be measured. In that case, definition of the property to be measured is not of importance to the so-called consensus validity theory. The *Standards* do not attend to this – the *Standards* claim that validating a test must begin with specifying a "construct the test is designed to measure" and a construct is a "concept or characteristic" (*Standards*, 11). Note, characteristics of a person and concepts are different things. Concepts being elements of thought and language, perhaps even ideas, are not something that we can measure whereas characteristics are (Maraun & Gabriel, 2013). For instance, the concept of a dog might encapsulate all sorts of things about dogs and the ways they might be related to each other (types of dogs, things dogs typically do, domesticated dogs on the couch, etc.) but a characteristic of a dog is something more precise and measurable – for now, we will use the term property. For instance, a dog's height and length would be characteristics of dogs – properties of dogs. Alternatively, dog breeds might have certain characteristics like "not enjoying getting their paws wet." In this way, the *Standards* are also incoherent. One example of a measurable construct that they provide is "mathematics achievement" – something that is not necessarily requiring of a measurement claim since mathematics achievement could just be whether a student performs what they are supposed to perform in a math class – this is not a mental or psychological attribute but a behavior of some sort. Instead of clarifying, the *Standards* move towards describing how test scores could be used.

In other words, there is no guidance from the *Standards* about how to define what one wants to measure and whether measurement is happening at all – though, scores are spoken of as if they are a measurement. Therefore, the *Standards* provide no guidance about construct definition. This is not to say that the *Standards* are of no use, but, at least in this case, they are not of much help.[9]

The position taken in this dissertation aligns with Newton and Shaw (2015)who argue that there is in fact no real consensus over the "best way to use the word validity" (Newton and Shaw, 2015, p. 183; as evidenced in their article by the immense number of different definitions of the word *validity* given by textbooks and educational and psychological measurement scholars). They maintain, though, that if we are to use the word validity, we should come to some consensus definition that is indeed technically sound, or, if not drop it all together. In fact, they are led to reject partially the argument that a consensus may be possible, invoking Wittgenstein's conception of a family resemblance, though, perhaps leaving out a very important aspect of Wittgenstein's writings on language games.[10] The idea of family resemblance is that, when using words and trying to parse meanings, "we see a complicated network of similarities overlapping and criss-crossing: similarities in the large and in the small" (Wittgenstein, 2010, sec. 66). These are called *family resemblances*. We also play language games that allow us to speak to each other and understand without

[9] This lack of attention to terminology ends up hamstringing the *Standards's* position on fairness, which, essentially, equates fairness to validity. By focusing on score interpretations and building evidence, the *Standards* make validity, to a broad extent, a matter of empirical investigation. Meanwhile, fairness is value laden. Yet, the selection of constructs and what they are, may indeed be very well improved by considering ethics (Borsboom & Wijsen, 2017).

[10] Wittgenstein, in discussing meaning said of some terms for which we play language games, said that "there is: The tendency to look for something in common to all the entities to which we subsume under a general term…Games form a *family* the members of which have family likeness. Some of them have the same nose, others the same eyebrows, and others again the same way of walking; and these likenesses overlap" (Wittgenstein, 2009, p. 106)(Wittgenstein, 2009, p. 106)(Wittgenstein, 2009, p. 106)

needing precision in every conversation, since, according to certain rules (the language game), grammar is used correctly or incorrectly based on context. The rules of one language game can be misapplied in a given context when different rules are needed. For instance, using the term *biker* to refer to cyclists of the Tour de France variety while speaking to the Hell's Angels could lead to some confusion. So, in essence, a family resemblance allows us to communicate because word meanings for people playing the same language game are close enough given family resemblances. However, in the case of the term *validity*, it is clear the rules of certain language games may be misapplied when there are completely different meanings of the term *validity*. There are some family resemblances – *validity* has something to do with tests, often in education or certification contexts – but beyond that, one may take the position of Borsboom and colleagues (a causal theory of test validity focused on measurement) or that of Kane and the *Standards* focused on justifying test score use in a given context (Kane, 2013) and neither would be wrong. I take the position that a dissolution of this conflict is necessary – the term validity will be avoided in this dissertation because it may cause more confusion.

Additionally, this confusion might be a function of the fact that most discussions that debate validity or invoke the notion of validity do not need the word validity at all! For instance, the statement, "I have validly measured" can be replaced by "I have measured" and "the instrument scores have a valid use" can be replaced by "instrument scores have a use" The case where we say "valid use," we want the word "valid" to do some extra work such as, <the test serves ethical and important ends>. Again, this does not need the word *validity*.[11] If one does not need the word validity to do any work in a sentence, this redundancy likely

[11] Andy Maul proposed this as a deflationary theory of validity.

violates some conversational maxims – for instance, raising the expectation of a listener to attribute extra meaning to the word validity that it does not import.

Newton and Shaw (2015) say that if we are to drop the term *validity,* we should have a technical or precise enough vocabulary to replace the term. Ironically, part of the problem with the term *validity* is that it refers to nothing since we do not have a precise enough vocabulary, as discussed above. What we want to head towards is a language and structure for why we should trust claims about a test (or test score). This language may already exist in the field of metrology, a formal study of measurement that also sets measurement standards in the physical sciences and engineering. Therefore, it is hoped that working toward ideals from metrology will help with clarity in psychological measurement claims - only deviating when social or political issues need be discussed. However, while metrology gives a place for property definitions, it does not help with *how to define* either - just what might be necessary.

One possible solution to the problem is an essential reorganizing of the *Standards*. For instance, a focus could instead be on generating key terms that cut across statistics and measurement. A brief attempt at this effort is introduced in chapter 3 of this dissertation. Instead of starting with notion of validity, one could start with the notion of how scores can be generated, and how they relate to measurement, necessitating a discussion of what is measured. Finally, instead of *validity* being a key term in the *Standards,* it could be considered one term of many, since there are so many ways that one may want to use a test and selecting among important cross cutting measurement terms may be more useful.

2.3.2 The psychometric scaling literature

As the passage from Borsboom (2006) that opened this dissertation indicated, what is measured needs be at the forefront of any communication of measurement results. However, it is not common practice in psychometrics. Much attention in psychometrics has been devoted to scaling. An example of scaling in the physical sciences would include generating a numerical scale for the expansion of mercury in a glass tube to measure temperature or how scores (or ability estimates) should be attributed to test takers. However, it is the position that of this dissertation that this is not sufficient for measurement. Following Rozeboom (1966), the notion of scaling is certainly important for measurement, but "'measurement' in the tough sense of the word must be distinguished from scaling" (Rozeboom, 1966, p. 170). One end of scaling can be thought of from the perspective of S.S. Stevens. Though Stevens provides a definition of measurement (provided below) that says measurement is only the assignment of numbers (or scale values) according to some set of rules (e.g. – a question on a survey with four response options may have each response option coded as 1-4, thus having it be said that this is measurement), there is no reference to *what* is measured. An alternative view of scaling may look toward the works of Lord & Novick or more recent technical literature on Classical Test Theory (CTT) which focuses on sums scored items and focuses on reliability or Item Response Theory (IRT) which focuses on scaling so that a latent ability of persons is posited and estimated from the statistical model. IRT models, being unidentified without some arbitrary constraints, are only scaled, for instance, when the distribution of person abilities or the distribution of difficulties are constrained to have a mean of 0 (there are other possible constraints to identify models see for instance, de Ayala, 2009; Feuerstahler & Wilson, 2019).

Scaling in psychometrics has been discussed without any reference to the attribute and only in technical terms. As such, while in principle one *could* make scaling about the attribute one hopes to measure where a unit is selected based on substantive concerns

(Briggs, 2019), the technical concerns of IRT have little to do with defining what one hopes to measure. As an example, we could combine a series of test questions ranging from calculus questions, Spanish language questions, and reading passage questions in English, and constrain a person distribution to have a mean of zero to identify a 1 parameter logistic item response model (1PL IRT) so each student has an estimated ability. In this case, it would be hard to interpret what is being measured – an ability to do what? Certainly, one could argue that there is something being measured, but that something would only become clear after identifying what it is and this only occurs outside of scaling.

This is not to say that if one uses an IRT model (or any reflective latent variable model for that matter), there are no ontological commitments. Ontological commitments are commitments to what is being measured – or in the gaudy words of Quine's (1948) title for his own essay on the matter, it's a consideration "On what there is." In the case of an item response model of the form,

$$P(x = 1 \mid \theta_s, \delta_i) = \frac{\exp[\alpha_i(\theta_s - \delta_i)]}{1 + \exp[\alpha_i \, (\theta_s - \, \delta_i)]}$$

which is the probability of answering an item, x correctly (coded as 1) conditional on the ability, θ, of student *s*, and difficulty, δ, of item *i*. The interpretation of the α parameter is up for grabs ontologically. In some sense it is a slope or discrimination parameter denoting the relationship, or correlation, between an item *i* and the attribute of interest or even as some denoting of a unit (e.g. de Ayala, 2009; Humphry, 2011). The most basic plausible idea is that this is a model *of* the data where the α parameter is effectively a way to model all influences on how the data came to be that are not the property (modeled as θ) of interest. When slopes vary by item, this is known as a 2PL IRT model (Lord & Novick, 1968;

different cumulative distribution functions or link functions can be used – typically they are normal or logistic cumulative distribution functions). When slopes are set equal to 1 across all items, this is sometimes known as the 1PL model, or in other settings, the Rasch model (Rasch, 1960/1980). Typically, the ability parameter, θ is treated as a random latent variable. The ontological commitment in this context is what θ corresponds to in the formula when the model is used for estimating an individual student's ability value. When this particular statistical model is applied for a given use, it represents, to some extent, how data would come to be if the model was right in the way it was supposed to be correct.

In the classic text about linking and equating which is effectively a matter of estimating person abilities (or test scores) across test administrations being on the same scale, Kolen and Brennan (2014), use converting between the Celsius and Farenheit temperature scales as an example of equating values. However, they say, this is not the same as typical issues in equating in educational testing. For instance, Kolen and Brennan (2014) say in the temperature context, "the relationship between the two scales is predefined" (p. 487) and if there are problems with temperature measurement conforming "exactly to the stated relationship, then there must be errors in measurements because the "construct" [quotations their own] that we call temperature is exactly the same for both scales" (p. 487). Alternatively, they maintain, this cannot be so in social science because tests "almost always measure at least some different constructs even if they have similar names" (p. 488). These statements are somewhat problematic owing to misconceptions about what leads to measurement results and the structure of measurement results more broadly (Mari et al., 2019; Maul et al., 2018a). However, it is worth acknowledging that the authors are saying that linking and equating have no use when there are different properties being measured.

Linking and equating have nothing to say about what is measured and *presuppose* that the same thing is being measured.

The Kolen and Brennan statement makes clear that psychometrics does not have a useful language for discussing uncertainty about what is measured. Given the status of the Kolen and Brennan (2014) text as foundational in the field of educational testing and speaking to larger issues about factor indeterminacy and scale indeterminacy in factor analysis and item response theory, it seems like a good example to use. When they say that educational tests measure multiple “constructs” (presumably, properties) and instruments measuring temperature do not, this is not quite correct. Instead, this is likely an expression of having more, what metrology might call, definitional and instrumental uncertainty in the realm of educational testing than in the realm of temperature measurement. As they say, educational tests measure other constructs and temperature instruments do not. However, broken down, this is an absurd statement. When Kolen and Brennan speak of “measurement error” they are referring to random errors but then they say measuring “other constructs” is not a form of measurement error. This is a bit confusing since measurement error, in a true value sense, would seem to be any measurement indication that deviates from the *true* value of what we intend to measure, but does not have any relevance to uncertainty due to the definition of the measurand. I assume this has to do with some mix of realism and CTT based theories, and not a concept of what measurement may be, nor is it an accurate account of what many historians, measurement scholars, and physical scientists think of measurement. The field of metrology has broken down types of error into different sorts of uncertainty (Giordani & Mari, 2012; Joint Committee for Guides in Metrology (JCGM) et al., 2008; Rigdon et al., 2019).

2.3.3 Metrological thinking

Broadly speaking, metrology, not unlike statistics, will speak of uncertainty about a measurement result in terms of random and systematic error. Random error is due to perturbations or random variation in measurement settings. Systematic error are errors that can be corrected for since they're known as always present in a measurement setting – perhaps due to external features influencing the measurement setting. That is, measurement instruments (like a thermometer) "act as a selector, interacting with the object under measurement with respect to a given quantity, the measurand" (Giordani & Mari, 2012, p. 5) where the measurand is what we would like to measure. Metrology defines so-called influence quantities – those things that are not the intended target of measurement, but the measuring instrument interacts with and change its reading. When Kolen and Brennan (or others) say a test or assessment does not measure the construct of interest but measures other things (or many things), I can only take this as an expression of uncertainty about *what* it is that is being measured and that there is a lack of confidence about the primary source of variation in scores in an uncontrolled way.

The aims in the field of metrology are to learn how to quantify this uncertainty or implement corrections for things that influence measurement results. This would be the case in a setting such as measuring temperature where barometric pressure interacts with the thermometer to change the temperature reading. This does not mean that the temperature is different. One can hopefully see that discussing thermometers as *not* measuring different things and educational and psychological tests as *indeed* measuring different things needs clarification. Namely, when we say something like a test in education and psychology measures different things, we seem to really mean that there is unaccounted for systematic measurement error due to unknown influence quantities, and this leads to a lot of uncertainty in test or survey scoring. However, in the case of temperature measurement, the sources of uncertainty, due to accepted thermodynamic laws, are relatively few (and, we think, mostly known). In the latter case, there is a model-based account of what is happening in the measurement process.

2.4 So, what are we working toward?

Why do we need to work on definition of attributes so carefully for the purpose of measurement? A brief aside on measurement itself might be useful with special attention paid to the philosophical deviations of measurement in the human sciences and psychological sciences. The history of measurement in psychology is not the explicit focus in this paper, and for that, one may want to turn to Michell (1997). However, contextual footing is necessary. Maul, Mari, Irribarra, and Wilson (2018) note that working society, engineering, and physical sciences puts a lot of faith in measurement results. It would be a strange world were we to not trust measurement results that have led to the construction of physical structures. Measurement has thus been clearly of importance to education and psychology as measurement may be perceived as granting legitimacy to a science. In the 1930s the British Association for the Advancement of Science organized a committee– the Ferguson Committee – to contest the feasibility of measurement in psychology (at the time, psychophysics – see, McGgrane, 2015; Michell, 1999). According to Stevens, “the committee was instructed to consider and report upon the possibility of `quantitative estimates of sensory events` – meaning simply: Is it possible to measure human sensation?” (Stevens, 1946, p. 677). Stevens claimed that what was really of discussion was a definition of measurement. He attempted to answer this question, crediting Norman Campbell – “we may say that measurement in the broadest sense, is defined as the assignment of numerals to objects or events according to rules." (Stevens, 1946, p. 677). Stevens then exposited the acceptable statistical techniques based on measurement scale types typically now learned in social science courses (nominal, ordinal, interval, ratio). Ironically, Zumbo and Kroc (2019) showed that these acceptable statistical techniques according to Stevens are not even consistent, using the example of computing covariances of nominally scaled variables when those variables are used as predictors in linear regressions. This should not be acceptable according to Steven's admissible statistical operations, but is in fact quite necessary and appropriate.

While psychology followed the tradition of operationalism and the scales of measurement of Stevens, McGrane (2015) notes that physical sciences and engineering turned toward the field of metrology to define and ensure the quality of measurement results.[12] In the long run, this has led to the development of the SI units (this discussion is excluding for now an additional direction, axiomatic and representational measurement theories). Stevens emphasized in his commitment to operationalism, that measurement was merely about the assignments and rules – thus less worry about what was being measured or a connection between measurement units and physical instantiation of those units.[13] The problem with Stevens' definition is that it was so broad as to be nearly useless in its over-permissiveness – any situation where numbers are assigned according to rules would be measurement (e.g. – Borsboom uses the example of the Dewey Decimal System, but one

[12] This is not a claim that the physical sciences are real sciences and human sciences are not or could not be (e.g. demarcating).

[13] In fairness to Stevens, he might debate this claim, as he said: "scales 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 numeral series." Though, McGrane (2015) notes, Stevens was most committed to consistent application of rules as opposed to thinking about properties that are quantitative and measurable.

might also see that player numbers in sports would fit Stevens's definition). The rules are not constraining nor descriptive enough to say what measurement is or what should plausibly accepted as a measurement result. Thus, others, including Michell (1997) and Maul et al. (2018) continue a quest in defining.

Maul et al. (2018; also see, Mari et al., 2021) ask, essentially, what leads to measurement results that are trusted? Phrased another way, they posit that measurement results are trusted, so what is it about the structure of a measurement process that leads to trust? What makes measurement *measurement*, as opposed to testing, assessment, or other forms of evaluation? In posing this question, they identify that trusted measurement results meet criteria of objectivity and intersubjectivity. One might see this as analogous to causal inference work, in which the veracity or trust in the causal claims or estimated treatment effects are rooted in the research design and not merely the statistical methods once data are collected (Morgan & Winship, 2015, chapters 1 & 2) . Maul et al. identify two basic features of trusted measurement results: objectivity and intersubjectivity. Objectivity is the extent to which values conveyed from a measuring instrument correspond to the intended attribute to be measured (e.g., length, "reading ability," temperature, and human resilience) and not other properties. Intersubjectivity is the extent to which measurement results can be interpreted in the same way "by different persons in different places and times" (Maul et al, 2018 p. 116). One can see that intersubjectivity is dependent on objectivity of the measuring instrument. It is perhaps no surprise that work from causal inference may serve as a nice analogy – Maul et al. and others (Markus and Borsboom, 2013), see measurement as a causal process in which a measuring instrument interacts directly or indirectly with the property under measurement. This property under study changes the indication of the measured attributes value. This can be within object (a child’s height measured two years apart) or between objects (person 1’s height vs person 2’s height at time 1) or both within and between – either will cause a change in indication value if the values within or between person are not the same, a counterfactual of sorts. The implication of this is that things that a measurement instrument is sensitive to have a causal role in the system.

However, Maul et al. (2018) name an important caveat. Objectivity requires a well-defined, specific definition of what is to be measured. I should be able to interpret measurement results later in the day or two days from an a measurement, even if the value of the measurement result of the particular object has changed in that time period, and there should be limited differences in the results of the width measurement depending on which instrument I use to measure that length. One can see that, also, by specifying the property under measurement, a measurer delimits what is not intended to be measured. Mari et al. (2021) and Maul et al. (2021) turn to the field of metrology for guidance instead of the field of psychometrics. Part of the reason for this is one of pragmatics. Metrology is an older and more principled way of studying measurement and measurement quality that has seen success across many fields of study. It is hard to say the same about psychometrics. Additionally, metrology focuses on measurement specifically, when this is not the focus of much of psychometrics, which, at different times, has been concerned about validity of inferences or statistics alone (though, the language of metaphor is certainly pervasive).

Metrology, in general, has more history in standardizing measurement practice. The Diplomatic Conference of the Metre in 1875 which established the SI (e.g, see, Quinn, 2011, 2017) also founded the International Bureau of Weights and Measures (BIPM). As Quinn (2011) documents, by 1791, there were already efforts to abstract the notion of length in terms of a single unit – the meter – as defined by swings of a pendulum at a given latitude for a given time. Instead, it was decided that the meter was to be defined via physical quantity – a platinum bar "containing small amounts of rhodium and iridium, 25 mm wide, 4 mm thick whose distance between the flat ends was defined as one metre" (Quinn, 2011, p. 9). National laboratories across the world are now signatories of the Diplomatic Conference of the Meter (Participating Laboratories - BIPM, n.d.), allowing for worldwide coordination of measurements as well as agreement and maintenance of standard units for measuring, for instance, time, length, distance, temperature, pressure, brightness of lights and electrical current. Since the founding, of the BIPM, units have been continually abstracted moving from physical idealizations (for example, the meter bar) to physical constants. As noted by Mcgrane (2015) citing Tal (2016), the second or unit of measurement for time is defined theoretically: "the invariant frequency of radiation emission by a caesium atom during a particular physical transition" (Mcgrane, 2015, p. 5). Today, among other documentation and purposes, the BIPM produces the International Vocabulary of Metrology (VIM) and the Guide to the Expression of Uncertainty in Measurement (GUM) which direct measurement efforts for science and engineering by taking advantage of this long history.

By contrast, *The Standards for Educational and Psychological Testing*, have a shorter history and much less indication of actual success or coordination. The *Standards* provide mostly direction in terms of validity theory, often conflating validity theory with measurement. While metrology has focused on developing the SI, sources of uncertainty, and definitions of units in terms of "physical theory" (Mcgrane, 2015, p.5), the *Standards* have focused on methods for reporting and using test scores based on general psychometric models with no focus on property definitions. The *Standards* primarily focus on statistical techniques for analyzing test score data – and not what those data relate to. One can express sympathy for the *Standards* though – educational and psychological measurement has not concerned itself, as noted by Borsboom above, with the same efforts for defining what is to be measured and how those measurements are realized, so there should be little consensus about how to define. Definitional work has provided little guidance for the *Standards* or anybody working in this realm – and the *Standards* are not an international collaboration among research institutes but instead consensus attempts among individual scholars and testing companies. So how would we (or should we?) make the *Standards* more like the VIM or the GUM? It is likely that a first step would be to focus on what is being measured and definition of those being measured. It is also possible that not everything should be admitted as measurable in a particular way as in the GUM, in which case the *Standards* should not be like the VIM or GUM. Questions about “what is measured” or “is it measurable?” are not an emphasis of the *Standards.* In other words, education and psychology are stuck somewhere between the *Standards* and metrology. This extremely brief sketch above may provide some sort of aspirational guidance. Of course, some might argue that definitional work is just not necessary.

2.5 Justification for Definitional Work in Measurement

While it might seem trivial that worrying about a coherent definition of what you intend to measure is important, others might worry that definitional work either has no obvious resolution, so is not worth considering, or that empirical means (such as statistical analyses) are the only ways we make progress in definitional work. In the realm of statistics, psychometrics, or measurement in human sciences broadly, some might argue that the application of enough statistics will lead to the correct definition. However, the notion of a correct definition is, to some extent, a social act – an act of naming situated in rules. The correct linguistic usage of a concept, where in our case, a concept might be a general idea of what is hoped to be measured, requires one "to employ it in accord with the linguistic rules that fix its sense, and to recognize an incorrect employment…is to recognize a departure from this normative employment" (Maraun & Peters, 2010, p. 128). In other words, the correct use of a concept, or in our case, the name of a concept, does not only come from merely empirical activities, but from norms, values, and histories. The process of *naming* an entity is not just an empirical one. In related papers, Slaney & Maraun (2007) and Maraun & Peters (2010) discuss how scientific practice is both conceptual and empirical. When discussing correct definitions or employment of concepts such as *resilience* or *literacy* one might have to venture into two related but separate enterprises:

> "Issues pertaining to the correct employments of denotative concepts are conceptual issues, whereas the study of the referents of these concepts, their causes, correlates, and properties are empirical issues" (Maraun & Peters, 2010, p. 128).

From the perspective of using *only* empirical means alone to identify or *discover* new entities, there is both philosophical and empirical work describing why this is impossible. To begin with the empirical, there are a number of articles demonstrating that even when statistical models fit data perfectly (given certain statistical criteria of fit), this is no guarantee that this model mimics the data generating process. For instance, it is extremely common to follow in Spearman's (1904) lead and use what are now a general suite of latent variable models such as factor models or item response theory model (IRT). This commitment to latent factor models in which some attributes cause responses to items is often justified by the presence of measurement error amongst item responses – the latent variables models are used to model shared variance among items while also modeling the unique variances conceptualized as so-called error – is noted by Rhemtulla et al. (2020). However, Rhemtulla et al. go on to show that this assumption, is, to some extent, overused.

In realms of educational and psychological measurement, the attribute targeted for measurement lead to decisions about people or groups of people. From a research perspective, in the words of Slaney and Maraun (2007), "science is ideally suited to resolving empirical uncertainties, it does not thrive in the company of conceptual confusions." (p. 107). However, I do not quite take this stance in this dissertation. Indeed, to do meaningful research, concern with the ontology of what is being measured requires that one does conceptual and coherence analysis (a view perhaps most coherent with analytic philosophy). If something does not make sense, either in accordance with what is already known or in accordance with the rules of grammar, or there is some circularity in reasoning, conceptual analysis will do better service for research than doing statistics. However, it is also going to be the standpoint of this dissertation that empirical work and attempts to construct measuring instruments also help with considering the nature of what is being measured. Examples of iterating between empirical work and conceptual work are described in Wilson (2005) as a process called construct mapping in which idealized levels of the attribute of interest are specified and observations that may serve as evidence of those levels are hypothesized. Attempts to elicit those observations are made and typically statistical tests are used to see if, for instance, the hypothesized ordering of the observations holds. In this way, exploratory, qualitative, or model testing may help resolve *certain* conceptual debates.

As later Wittgenstein might have noted, words direct our thinking, our conceptualizations, and the meanings of words are intimately tied to the use of those words (this may also be a Vygotskian notion, e.g. Vygotsky, 2019). In other words, pushes to formalize will be shaped by starting points which, effectively, are the common language terms we use. Concepts and what we may call facts in our vernacular or values (or evidence of a fact) may not be so clearly delineated. The distinction between observational and theoretical terms, a source of classic debate in philosophy of science will be shown to be not such a clear distinction – meaning definitions cannot rest on this distinction or distinctions such as "latent variable" vs "observed variable." Some of the recent calls to formalize psychological *theories* are about increasing transparency. Likewise, it is imperative that, in the words of Michael Billig, (2011):

> We should be looking carefully at the ways that people use language to do what they do. We should not assume that the 'real' objects of psychological inquiry lie behind these familiar words – as if we should be searching for, and naming, 'the real things' of which people are unaware and which cause them to do what they do. (Billig, 2011, p.6).

Perhaps another way to say this is that one should be wary of technical language standing in for saying anything. Billig (2011) and use-based advocates of word meaning stand in partial contrast to the claim by Paul Meehl (1992) that defining a psychological concept is somewhat of an empirical matter that "depends on how the world is" and that the word used is unimportant (Meehl, 1992, p. 120). If Meehl (1992) were to include language use as part of the way the world is, perhaps this position would be more defensible. However, both Billig and Meehl may agree that what is important is the concept that is demarcated by the word. Meehl is perhaps missing a sense that Billig emphasizes from Wittgenstein in which the language used is important as it gives rise somewhat to what is believed to exist – that the extra meaning it comes with may provide a window into what is meant by the user of the words, which includes oneself as the user. As Billig states, sometimes formalization (left undefined), is the wrong task, and technical terms "can give the appearance of being technically precise, whilst, in point of fact, being highly imprecise" (Billig, 2011, p. 13). In other words, using active verbs might have been more precise than appending a name to the process as if the name was given absent human beings (e.g. – calling something, *resilience* without a history of how that term came to be). The conundrum thus presented is that while communication requires shared meaning, an appeal to technical language or a specific vocabulary that should nail down a shared meaning does not guarantee that terms share meaning without referencing something specific (a classic example in educational assessment would be the appending of *ability* to any attribute that is supposed to vary – "reading ability").

2.6 Historical Considerations and Discussion

Debates about what is to be measured may seem isolated to ivory towers, and maybe so. Nonetheless, these debates influence policy decisions about tests, and hence, types of decisions to be made about students. A brief historical account of the rise of educational testing in the United States helps show that educational testing and psychological assessment are not just a matter of research but also a matter of political power. Grouping certain forms of educational research as a special case of psychological research because both often deal with concepts of the mind, Teo (2018) questions what the very subject matter of psychology is to create something that is uniquely called *psychology*. Handing authority to researchers via terms like *measurement*, *expert*, or, *science,* leads to the setting that "what psychologists define, research, and conclude contributes, once disseminated to the public, to the co-

construction of the very identities of subjects, which cannot be conceptualized adequately without a concept of power" (Teo, 2018, p. 30). Given the subject-focused nature of psychology and the way it influences how people lead their lives, definitional work for the purpose of science is an epistemological and ethical issue – making it about "scientific correctness" (Teo, 2010, p. 298).

As documented by Reese (2013) the early history of educational assessment in the United States was explicitly a political move by Horace Mann and colleagues to oust headmasters (or grammar masters) in Boston area schools, with first (surprise/unannounced) written tests given in 1845. Reese (2013) described the setting of these surprise exams:

> "Battles between Mann and the grammar masters became personal, professional, and political, lethal combinations that shaped every aspect of the 1845 exams" (Reese, 2013, p. 59)

According to Reese (2013), Mann - authorized by local Boston school committee officials and taking advantage of, at times, mixed public sentiments about the quality of Boston schools (comparing Boston schools' evaluation methods [14] with supposedly objective methods used abroad for assessing teacher, headmaster, or school quality) - wielded surprise *written* tests as a means to show the inadequacy of the schools and headmasters. Prior to these written tests, schools held ceremonies in which recitals, or so-called emulation ceremonies, were presented to the local community. Star pupils were selected to recite answers as they were grilled by headmasters. Memorized student answers were sources of pride and led to communities bestowing adulation upon (male) school headmasters. Reese's

[14] Reese (2013) described so-called emulation in which star pupils were selected to recite (likely rehearsed) answers as they were grilled with questions by school headmasters .

historical telling describes how the move from single classroom settings in U.S. education prior to 1845 to age-grading and standardization for the sake of comparison followed in less than a few decades. Reese describes San Francisco school board members less than a century later:

> "They advocated the usual innovations endorsed by northern reformers: better age-graded classrooms, uniform textbooks, women teachers for the lowest grades, and the establishment of a high school (founded as the Union Grammar School in 1856)" (Reese, 2013, p. 162).

This brief historical summary shows how early tests in the United States served a very specific purpose to oust a group of people and include a lower paid group of instructors by changing, in effect, what was assessed - moving from memorizing answers and presenting orally to students answering written questions for which they were not prepared. One should be aware of this history – tests are not a-historical. Of course, the great irony according to Reese's history is that *written* standardized testing is often now accused of promoting students memorizing answers and learning how to take tests or so-called *teaching to the tests* – exactly what tests were meant to combat. Choosing what and how to measure, test, or assess, especially in the realms of public schools can be seen as a political act not without consequences. This is, perhaps, the worry around debates involving NAEP reading tests or similar district, state, or national level work (or student groupings often made along institutionally invoked racial or ethnic categories). In other words, what is decided to be measured is an invocation of what is relevant – and not necessarily a change in the concept of the thing that is measured – though, certainly, that is also possible.

For the researcher either crafting instruments based on common conception in the literature around a given area of study or the analyst working with data from these instruments taking heed of lessons from Arendt's 1963 book - *Eichmann in Jerusalem: A Report on the Banality of Evil* would be wise. The book is Arendt's adaptation from her reporting for the *New Yorker* on Adolf Eichmann's tribunal in a Jerusalem District Court in 1961 following his capture in Argentina. Eichmann was, in essence, a Nazi bureaucrat managing the deportation of Jews to other countries, ghettos, and death camps – he's often considered a member of the Nazi party key to carrying out the Holocaust - answering a call to carry out what the Nazis called, "The Final Solution to the Jewish Question" (see, for instance, Browning, 2004). Arendt's report and accompanying analysis effectively documented how Eichmann was but merely following orders, looking after his career. While Arendt claims that Eichmann was certainly not completely unaware of what he was doing, she argues that what allowed him to carry out his order to work toward extermination of Jews was ignoring implications of his actions in the name of following directions, doing the paperwork. In the postscript to her book, Arendt says of Eichmann:

> "He was not stupid. It was sheer thoughtlessness - something by no means identical with stupidity – that predisposed him to become one of the greatest criminals of that period… That such remoteness from reality and such thoughtlessness can wreak more havoc than all the evil instincts taken together which, perhaps, are inherent in man – that was, in fact, the lesson one could learn in Jerusalem." (Arendt, 1964, p. 134)

Arendt claims that Eichmann, aside from his own career had no motives. In this way, without comparing the practice of psychometrics, measure construction, or reliance on pre-defined, operational definitions to the evils of Nazism, the lesson Arendt hopes to present is that thoughtlessness (in our case) about bureaucratic practice - psychometrics, educational or psychological measurement, can have drastic consequences. Now, to be fair, though unsurprising, tapes from 1957 were recently unveiled that recorded Eichmann being very aware and even proud of his work, somewhat invalidating Arendt's basis of the view, or at least, partially (Kershner, 2022). However, the lesson may still hold or still stand, especially for statisticians. Being merely an analyst, methodologist, or defining only based on the literature is not absolving – the lesson from Arendt being that we owe thoughtfulness to those who are measured. Or in the words of Borsboom & Wijsen (2017) we cannot treat social consequences as easily discoverable, "as if educational tests and testing agencies can generally be expected to end up on the good side of history" (p.440). It is thus an ethical imperative to consider what we are measuring, sometimes requiring not a merely empirical, results oriented view – but trying to consider questions such as – is this the right thing to do?

How can we begin to pay attention to these sorts of things to keep a skeptical mind? Here I propose as starting place, a combination of conceptual analysis of the definitions of the properties of people we try to measure, understanding their coherence and potential empirical tools to understand how to iterate and improve these definitions. Considerations of what we might be doing when we define – how a definition, which is inherently stipulative to a degree – is considered as well. In the following chapter, I aim to introduce some initial tools for refining or outright starting anew, definitions of properties in the social sciences.

References

American Educational Research Association, American Psychological Association, & and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Arendt, H. (1964). Eichmann in Jerusalem : a report on the banality of evil . In *Eichmann in Jerusalem : a report on the banality of evil* (Rev. and enl. ed.). Viking Press.

Armstrong, D. M. (1980). *Nominalism and Realism: Volume 1: Universals and Scientific Realism* (Vol. 1). Cambridge University Press. https://philpapers.org/rec/ARMNAR-3

Billig, M. (2011). Writing social psychology: Fictional things and unpopulated texts. *British Journal of Social Psychology*, *50*(1), 4–20. https://doi.org/10.1111/J.2044-8309.2010.02003.X

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory Construction Methodology: A Practical Framework for Building Theories in Psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Borsboom, D., & Wijsen, L. D. (2017). *Psychology's atomic bomb*. https://doi.org/10.1080/0969594X.2017.1333084

Briggs, D. C. (2019). Interpreting and visualizing the unit of measurement in the Rasch Model. *Measurement*, *146*, 961–971. https://doi.org/10.1016/J.MEASUREMENT.2019.07.035

Browning, C. R. (2004). The origins of the final solution: The evolution of Nazi Jewish policy, September 1939-March 1942. *The Origins of the Final Solution: The Evolution of Nazi Jewish Policy, September 1939-March 1942*, 1–615. https://doi.org/10.2307/20034240

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. Guilford Press.

Dunne, G. (2020). Epistemic injustice. *Encyclopedia of Educational Philosophy and Theory. Singapore: Springer Nature*, 2–7.

Feuerstahler, L., & Wilson, M. (2019). Scale Alignment in Between-Item Multidimensional Rasch Models. *Journal of Educational Measurement*, *56*(2), 280–301. https://doi.org/10.1111/JEDM.12209

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

Fried, E. I. (2021). Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature. *Https://Doi-Org.Proxy.Library.Ucsb.Edu:9443/10.1080/1047840X.2020.1853461*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

George, T. (2021). Hermeneutics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University.

Giordani, A., & Mari, L. (2012). Measurement, models, and uncertainty. *IEEE Transactions on Instrumentation and Measurement*, *61*(8), 2144–2152.

Hacking, I. (1999). The social construction of what? . In *The social construction of what?* Harvard University Press.

Humphry, S. M. (2011). The role of the unit in physics and psychometrics. *Measurement*, *9*(1), 1–24. https://doi.org/10.1080/15366367.2011.558442

Joint Committee for Guides in Metrology (JCGM), Jcgm, J. C. F. G. I. M., Bipm, Iec, Ifcc, Ilac, Iso, Iupac, Iupap, Oiml, Jcgm, J. C. F. G. I. M., Joint Committee for Guides in Metrology, Jcgm, J. C. F. G. I. M., Drive, B., Issue, F.--, Drive, B., Williams, T., Kelley, C., Campbell, J., … Parkway, W. M. (2008). Evaluation of measurement data — An introduction to the “Guide to the expression of uncertainty in measurement” and related documents. *International Organization for Standardization Geneva ISBN*, *3*(October).

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kershner, I. (2022, July 4). *Nazi Tapes Provide a Chilling Sequel to the Eichmann Trial - The New York Times*. The New York Times. https://www.nytimes.com/2022/07/04/world/middleeast/adolf-eichmann-documentary-israel.html

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363.

Kolen, M. J., & Brennan, R. L. (2014). Linking. *Test Equating, Scaling, and Linking*, 487–536. https://doi.org/10.1007/978-1-4939-0317-7_10

Ladson-Billings, G. (2006). From the Achievement Gap to the Education Debt: Understanding Achievement in U.S. Schools. *Educational Researcher*, *35*(7), 3–12. http://www.jstor.org/stable/3876731

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores. In *Statistical theories of mental test scores.* Addison-Wesley.

Maraun, M. D. (1998). Measurement as a Normative Practice. *Theory & Psychology*, *8*(4), 435–461. https://doi.org/10.1177/0959354398084001

Maraun, M. D., & Gabriel, S. M. (2013). Illegitimate concept equating in the partial fusion of construct validation theory and latent variable modeling. *New Ideas in Psychology*, *31*(1), 32–42. https://doi.org/10.1016/j.newideapsych.2011.02.006

Maraun, M. D., & Peters, J. (2010). What Does it mean than an Issue Is Conceptual in Nature? *Journal of Personality Assessment*, *85*(2), 128–133. https://doi.org/10.1207/s15327752jpa8502_04

Mari, L., Carbone, P., & Petri, D. (2012). Measurement Fundamentals: A Pragmatic View. *IEEE Transactions on Instrumentation and Measurement*, *61*(8), 2107–2115. https://doi.org/10.1109/TIM.2012.2193693

Mari, L., Maul, A., & Wilson, M. (2019). Can there be one meaning of “measurement” across the sciences? *Journal of Physics: Conference Series*, *1379*, 12022. https://doi.org/10.1088/1742-6596/1379/1/012022

Mari, L., Wilson, M., & Maul, A. (2021). *Measurement across the sciences: developing a shared concept system for measurement*. Springer.

Maul, A. (2013). On the ontology of psychological attributes. *Theory & Psychology*, *23*(6), 752–769. https://doi.org/10.1177/0959354313506273

Maul, A., Mari, L., Torres Irribarra, D., & Wilson, M. (2018b). The quality of measurement results in terms of the structural features of the measurement process. *Measurement*, *116*, 611–620. https://doi.org/10.1016/J.MEASUREMENT.2017.08.046

Mcgrane, J. A. (2015). *Stevens' forgotten crossroads: the divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century*. https://doi.org/10.3389/fpsyg.2015.00431

Meehl, P. E. (1992). Factors and Taxa, Traits and Types, Differences of Degree and Differences in Kind. In *Journal of Personality* (Vol. 60). https://meehl.dl.umn.edu/sites/g/files/pua1696/f/150factorsandtaxa.pdf

Michell, J. (1999). Measurement in Psychology: Critical History of a Methodological Concept. *Measurement in Psychology a Critical History of a Methodological Concept*. https://doi.org/10.1017/CBO9780511490040

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.

National Assessment Governing Board. (2021). Reading Framework for the 2026 National Assessment of Educational Progress. In *NAEP*.

Newton, P. E., & Shaw, S. D. (2015). Disagreement over the best way to use the word ‘validity’ and options for reaching consensus. *Https://Doi.Org/10.1080/0969594X.2015.1037241*, *23*(2), 178–197. https://doi.org/10.1080/0969594X.2015.1037241

Noorgard, K. (2008). Human Testing, the Eugenics Movement, and IRBs. *Nature Education*, *1*(1), 170. https://www.nature.com/scitable/topicpage/human-testing-the-eugenics-movement-and-irbs-724/

Orilia, F., & Paolini Paoletti, M. (2022). Properties. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022). Metaphysics Research Lab, Stanford University.

Participating laboratories - BIPM. (n.d.). Retrieved May 7, 2022, from https://www.bipm.org/en/cipm-mra/participation

Quine, W. v. (1948). On what there is. *The Review of Metaphysics*, *2*(1), 21–38.

Quinn, T. (2011). *From Artefacts to Atoms: The BIPM and the Search for Ultimate Measurement Standards*. Oxford University Press.

Quinn, T. (2017). From artefacts to atoms - A new SI for 2018 to be based on fundamental constants. *Studies in History and Philosophy of Science Part A*, *65–66*, 8–20. https://doi.org/10.1016/J.SHPSA.2017.07.003

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. The University of Chicago Press.

Reese, W. J. (2013). Testing wars in the public schools : a forgotten history . In *Testing wars in the public schools : a forgotten history*. Harvard University Press.

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1). https://doi.org/10.1037/met0000220

Rigdon, E. E., Becker, J.-M., & Sarstedt, M. (2019). Factor Indeterminacy as Metrological Uncertainty: Implications for Advancing Psychological Measurement. *Multivariate Behavioral Research*, *54*(3), 429–443. https://doi.org/10.1080/00273171.2018.1535420

Rozeboom, W. W. (1966). Scaling theory and the nature of measurement. *Synthese*, *16*(2), 170–233. https://doi.org/10.1007/BF00485356

Scheffler, I. (1963). The language of education. *Philosophy*, *38*(144).

Searle, J. R. (1995). *The construction of social reality*. Simon and Schuster.

Slaney, K. L., & Maraun, M. D. (2007). There are no “specific correct” usages of concepts, only correct usages: Linguistic rules and the bounds of sense. *Journal of Theoretical and Philosophical Psychology*, *27*(1), 104–112. https://doi.org/10.1037/h0091284

Spearman, C. (1904). &quot;General Intelligence,&quot; Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201. https://doi.org/10.2307/1412107

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*(2684).

Tal, E. (2016). Making Time: A Study in the Epistemology of Measurement. *Https://Doi-Org.Proxy.Library.Ucsb.Edu:9443/10.1093/Bjps/Axu037*, *67*(1), 297–335. https://doi.org/10.1093/BJPS/AXU037

Tal, E. (2019). Individuating quantities. *Philosophical Studies 2019 176:4*, *176*(4), 853–878. https://doi.org/10.1007/S11098-018-1216-2

Teo, T. (2010). What is epistemological violence in the empirical social sciences? *Social and Personality Psychology Compass*, *4*(5), 295–303.

Teo, T. (2018). *Outline of theoretical psychology*. Springer.

van Rooij, I., & Blokpoel, M. (2020). Formalizing Verbal Theories. *Https://Doi.Org/10.1027/1864-9335/A000428*, *51*(5), 285–298. https://doi.org/10.1027/1864-9335/A000428

Vygotsky, L. S. (2019). Mind in Society. In *Mind in Society*. https://doi.org/10.2307/j.ctvjf9vz4

Wartenberg, T. E. (1992). *Rethinking power*. SUNY Press.

Wijsen, L. D., & Borsboom, D. (2021). Perspectives on Psychometrics Interviews with 20 Past Psychometric Society Presidents. *Psychometrika*, *86*(1), 327. https://doi.org/10.1007/S11336-021-09752-7

Wijsen, L. D., Borsboom, D., & Alexandrova, A. (2022). Values in psychometrics. *Perspectives on Psychological Science*, *17*(3), 788–804.

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Lawrence Erlbaum Associates.

Wittgenstein, L. (2009). *Major Works: Lugwig Wittgensteing* (First). Harper Collins Publishers.

Wittgenstein, L. (2010). *Philosophical investigations*. John Wiley & Sons.

Zumbo, B. D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832

Zumbo, B. D., & Kroc, E. (2019). A Measurement Is a Choice and Stevens' Scales of Measurement Do Not Help Make It: A Response to Chalmers. *Educational and Psychological Measurement*, *79*(6), 1184–1197. https://doi.org/10.1177/0013164419844305

Chapter 3
Transparently defining properties for measurement in the human sciences

In this chapter I aim to introduce ideas from philosophy of language for fostering researcher coordination in education and psychology to continually improve measurement efforts and definitions of that which we aim to measure. To illustrate, I use the two cases of research areas already introduced in chapter 1. The quality and the consensus of definitions of properties in the areas of *resilience* scholarship and *literacy* scholarship will be shown to be in very different places depending on research area These differences can be enlightening for understanding more mature or productive definitional efforts. For instance, what differentiates definitional efforts for the sake of measurement in academic contexts like reading and literacy from definitional efforts in say, resilience? We may also want to define terms to refer to properties in general, absent any measurement efforts.

Attending to the above goals requires some conception of what we are doing with language when we define. Consider the activity of reading. There is often a desire to produce inferences about cognitive processes that are causally responsible for the act of reading (or understanding and comprehending; a classic account might be from Kintsch & van Dijk, 1978). Alternatively, for measurement in psychology (dealing with things not necessarily typically associated with academic or schooling contexts), there might be a desire to learn about mental or internal processes or states that seem to derive from personal experiences, feelings, and language.

In the previous chapters, the field of metrology and the VIM were introduced. A core concept in metrology is the measurand – that which we'd like to measure such as my height or a student's reading ability (reading comprehension ability of a given student is the measurand). However, we need a precise way of defining the measurand to characterize, for instance, how accurate measurement efforts are or, simply, to communicate about what is measured. The VIM, along with the *Guide to the Expression of Uncertainty in Measurement* (Joint Committee for Guides in Metrology (JCGM), 2008), also introduce the concepts of measurement uncertainty (though, with slightly different definitions between documents).

The GUM says that “uncertainty of the result of a measurement reflects the lack of exact knowledge of the value of the measurand” and that “the result of a measurement after correction for recognized systematic effects is still only an estimate of the value of the measurand because of the uncertainty arising from random effects and from imperfect correction of the result for systematic effects” (Joint Committee for Guides in Metrology (JCGM) 100:2008, 2008, section 3.3.1, p. 5). For the sake of clarity, we can colloquially think of uncertainty as “that which we do not know.”

I might take a series of measurements of my height and each measurement could lead to a slightly different measurement result. One can see why specifying uncertainty is important in measurement. Uncertainty provides room for doubt but is also important for considering the plausibility of different outcomes (I might say somewhat colloquially and in a Bayesian credible interval sense, that there is a 95% chance of my height being between 5’7” and 5’9”). If I have a lot of doubt, I might be more cautious in certain situations. For instance, if you administer a reading test to a student with a lot of uncertainty associated with a resulting score, how would you decide about the student’s reading comprehension ability? If you know that the measurement results are uncertain because of systematic effects related to something you do not intend to measure, we might try to adjust for that effect. However, this requires a fine or relatively precise conception of what we are trying to measure since definitions demarcate.

The GUM provides a list of numerous sources of uncertainty. One of these sources is “incomplete definition of the measurand” or “imperfect realization of the definition of the measurand” (Joint Committee for Guides in Metrology (JCGM), 2008, 3.3.2, p.6) among others. The GUM defines the objective of measurement around information gain about the measurand and “a measurement therefore begins with an appropriate specification of the measurand, the method of measurement, and the measurement procedure” (section 3 of the GUM, p. 4). The GUM argues that variation in measurement results even under the same conditions are the results of external influences that we do not want to measure – these are influence properties or quantities. Uncertainty is meant to be estimated via statistical means, for instance attributing shape and scale parameters to probability density functions (PDF) characterizing the uncertainty. Measurement uncertainty can also come from information related to background knowledge from previous measurement results. The problem, in the end, is that the GUM is most concerned with “the expression of uncertainty of a well-defined physical quantity – the measurand” (Joint Committee for Guides in Metrology (JCGM), 2008); it is hard to imagine psychological properties being well defined, at the moment.

Unfortunately, the VIM and the GUM, while providing some useful language for talking about measurement, measurement uncertainty, and what we would like to measure, do not provide guidance about how to define (potentially non-physical) properties. In the other sense, definitional uncertainty even in the *GUM* is salubrious because it implies doubts about definition are present even in well-defined settings. It also means that as we learn more, we can increase definitional uncertainty because we start knowing more about what we don’t know (Grégis, 2015). I would like to posit that sometimes definitional uncertainty in educational and psychological measurement is not always a matter of limits of what we can know, but a matter of trying to reach consensus about the coherence of ideas around a single property to measure.

In this chapter, I will work through definitional debates using two case studies. I aim to frame these debates from the perspective of philosophy of language and provide language and logic for thinking that might provide a useful way to move definitional discussions forward. I will then try to attempt to place these views in a philosophy of science – posing the question, "since we can imagine many things in the psychological realm, what should we admit into our ontology?" Paying attention to definition and language will provide a defense against over admission to this ontology. I aim to wrap up with a discussion on the perspective of latent variables and latent variable modeling and their referents in the real world. This will introduce language for the next chapter of this dissertation and how we will use the philosophy of science and metrological terminology to motivate statistical analyses that inform definitional efforts in reading comprehension

3.1 What Makes Something Measurable?

The goal of this chapter is to articulate definitions for measurable properties. The question about what makes something measurable is a large one and there are many deep resources for thorough discussions (e.g. Borsboom, 2005; Humphry, 2011; Mari et al., 2021; McGrane, 2015; Michell, 1997, 1999, 2005). However, to understand what makes definitions of properties to be measured fruitful for measurement, we need to understand what we might work towards. The question of measurability is in part ontological and in part epistemic. The epistemic part of measurement requires that we have some way of interacting with, knowing about, and modeling the thing we would like to measure. We measure properties of objects, so that means the property need be realizable when measuring. The definition of what we are measuring and its level of specificity is a function of how much we know about that property as well as the *needed* specificity or accuracy for a given purpose. Measurement is defined by the VIM as a "process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity" (JCGM, 2012; 2.1), placing human judgement as firmly part of measurement. This requires transparency about not just the data generation process but the ideation around, and knowledge of, the thing to be measured. For instance, if we say, “resilience is a latent variable” – this is not transparent because it tells us nothing about the property we would like to measure, how the idea of the property came to be, or how we might realize the property in a measurement process. By *realize the property*, I mean we have to have some way to interact with the property, or surmise that some observation is a particular instance of that property (e.g. reading comprehension ability is realized when a student reads a book and can provide a summary of the book’s argument even when those arguments are not made directly in text). However, it is also maintained that there is an ontological component of measurement. A measurement requires that we know and trust what is and is not the property of interest and its nature (Maul et al., 2018a; Tal, 2019). If we go through a measurement process that yields some strange result, we may reject that instance as a measurement because of knowledge about the way the world is.

Thus, we need to be able to identify the property’s structure, which brings about some ontological concerns. Steven’s (1946) argument that measurement “is defined as the assignment of numerals to objects or events according to rules” is rejected because it is overly permissive about rules being the justification for measurement since rules do not need to have any coherence or accordance with reality. However, we want other forms of justification, such as some representational correspondence (Mari, 2013). That is, when we say that student *a* has a higher reading ability than student *b*, there should be some correspondence between the notion of reading abilities reported in a measurement and the structure of reading ability in which a person can have a higher-valued reading ability. The challenge is justifying this structure. It is the position here, that indeed, a measurable property has a structure or realization that has some homomorphism – a mapping in which a measurement result resembles a characteristic of the object under measurement to the extent needed for the purpose of measurement. This implies also that properties have to be non-arbitrary universals. Universals can be realized, though some universals are dependent such that we might say reading ability is dependent on an object being a person or even another dependent universal such as *text decoding ability* (Arp et al., 2015).

To visualize how properties that we aim to define relate to measurement, figure 3.1, below, shows how we might conceptualize the measurement process when it involves a person taking a test or survey. Here, we have a multitude of person properties that are dependent on the person to which they belong. These properties are thought to cause a person to act or behave the way they do. The properties included are both the target property and other properties (e.g. a person's reading ability, attention span, and mathematics abilities). Items might be test questions or survey items. They have features such as use of particular words, formatting, ways in which responses are required (open ended or multiple choice), and mode of presentation (interview questions from an interviewer, paper and pencil tests, or computer-based). The person properties and item features interact and something leads to a person's response. This response will be probabilistic (ignoring, for the time being, why there is probability at all). Regardless, we see the centrality of the property of interest needing to be causally connected to the observed data. A manipulability theory of causation will be most relevant throughout this chapter. We shall conceive of relevant properties for measurement in terms of human agency and *potential* manipulation even if it requires a thought experiment where I may potentially intervene on the property (for example, had I improved my writing ability and clarity in thinking, this dissertation would be a lot shorter).

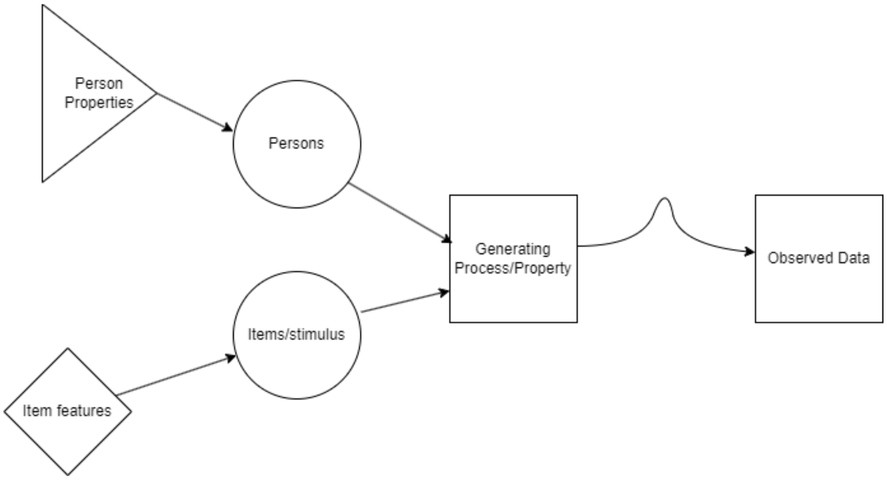


FIGURE 3.2 ***THE OBSERVED DATA GENERATING PROCESS WHEN A PERSON TAKES A TEST. THE CURVED LINE FROM THE PROPERTY INVOLVED IN THE GENERATING PROCESS TO THE OBSERVED DATA SYMBOLIZES THAT THIS IS NON-DETERMINISTIC.***

This is all to say, measurability of a property requires our own specification and knowledge of a property, ideas about its form and function, and a definition of the property that tells us what that property is. Some, like Tal (2019) or van Fraassen, will refer to this notion as the model dependence of measurement. Some guiding questions for considering the measurability of a property might be:

1. Do I know enough about the property to define and demarcate it from other properties? If not, why?
2. Do I know enough to define the property in a way that when we try to measure it, I know how the instrument and person will interact and what other properties are involved in the interaction of the person and instrument?
3. Do I know enough about the property that definitions might lead to a measuring instrument where results from the interaction of the person and the measuring instrument will be known to be about the property of interest (Mari et al., 2012;

objectivity) and the results can be interpreted in the same way across time and place (intersubjectivity; Mari et al. 2012; perhaps holding *certain* contexts constant)?

4. Finally, can I specify a cognitive model of the property or model of the measurand that specifies its algebraic structure? Do I think that it is quantitative in some regards or is it a categorically structured property (such that changes in property values are either quantitative in nature or qualitative)?

As noted, this means that measurement, its interpretation, and the things we measure are dependent on the models, assumptions, and *language* we use to describe those properties. In the words of van Fraassen (2012), whether something counts as measurement "is a question whose answer is in general determined by theory, not solely by operational or empirical characteristics" (p. 782). We can see that we have a model of a measurement process above in a generic testing situation about how properties interact and lead to a result and each of these properties may have a theory about them. We can say that this theory of measurement has a particular syntax – the (proposed) rules or structure - and a particular semantics, connecting them to a meaning.

Thus, we should think of definition as a tool in research just like any other tool. Ogden & Richards (1925) argue that definition should "direct travelers from all quarters to any desired point" (p. 116). This connects with a position on modeling where we see models as guiding tools. Giere (2009) argues that "agents (1) intend; (2) to use model, M; (3) to represent a part of the world, W; (4) for some purpose" (p. 274). We can look at definitions in the same way. In fact, models *can be* definitions or definitions can be viewed as models.

Just as statistical models are often used as one way of understanding a given research scenario, definitions are also tools for directing, coordinating, and allowing for correction. Debates about definitions in research or science are not meant to be about the word itself, but the property itself. Definitions, that is, do not create the referent of the word.

One way to understand the sorts of definitional challenges we face is through case studies or characterizations of definitional debates in specific domains. Here, I've included what I think is a relatively unproductive definitional debate where the nature of assigning a particular word meaning is the focus of the debate (research in human resilience) and one where the definitional debate is more about the edges of the idealized attribute (reading comprehension). These initial case studies are brief attempts at some linguistic and conceptual framing.

3.2 What can definitional analysis accomplish?

Why should scientists doing research or measurement be concerned with language? One answer is that we use words to think about or refer to entities in the world. For instance, we differentiate between math and reading using the words *math* and *reading* but also because we ask questions "what is reading comprehension?" The way we answer these questions will surely rely on a history of how these words are used and what they refer to. Simply, we use natural language to describe what we'd like to measure. For instance, terms like *length*, *width*, *reading comprehension ability,* or *resilience,* are terms that have varying levels of meaning specificity that we use to direct actions or investigation. In general, we are often worried about the correspondence of claims or terms to reality. These terms in natural language or some other form (e.g. statistical parameters in a model) are supposed to correspond to *something.* We often use language to understand correspondence between our ideas and the things those ideas, or concepts, might refer. Part of the challenge of defining an attribute to measure is that some model of the attribute need be constructed but the model is not the attribute, and there will always be omitted details. Brown articulates that "analysis of language provides insights into the nature of the cognitive processes used in reasoning and into the structured nature of our understanding of the physical world" (Brown, 2003, p. 5). We return then again to the notion of definitional work as a form of model building, below.

Grégis (2015) provides a nice example of measuring the length and width of a mostly rectangular table. In this case, the table is modeled as having a rectangular shape which is an idealization of similarity between the table and a rectangle. This allows us to speak about the table width and length (if the table had rounded corners, we will likely still treat it as a rectangle in most settings) based on a mutually understood idealization.

Definitional analysis for the sake of *measurement* aims for transparency. For instance, we may aim to move from defining "reading ability" as *whatever it is that causes a student to read* to trying to describe that causal force. The black box might contain what Kintsch and van Dijk call the micro level (even as low level as text decoding) or the macro level (piecing arguments together from within and outside the text) of reading comprehension. Definitional analysis for the sake of *measurement* aims for enough clarity to be fallible – for instance, by including ideas in a structured way such that observations might raise doubts about the world being this way (e.g. at least partially Popperian). Without transparency, it is not clear how a definition can be identified as incomplete (aside from strange use cases of words) a key principle of scientific research. Said another way, there is a belief in modeling that there are models of and models for (Giere, 2009; Gouvea & Passmore, 2017) and models (and measurement models) serve different based on various needs for accuracy but definitional analysis can help direct our actions in modeling. This is true of language as well, where words are spoken with a given intention. *Models of* refer to models that aim to represent. But this is not the only modeling concept. The extent to which this idealization occurs will be dictated by use, so this *model of* is also a *model for*. The *model for* conception will to some extent dictate the way we define something. It is argued here, then, that the purpose of measurement and definition of the measurand will be intertwined – and this tying together will help with iteration and improvement. And, indeed, we can use empirical means to decide on influences of the measurand that may help us better define it. For instance, Arya et al. (2011) found that *bacteria* is a word that many younger students knew the meaning of, even though, by some computational methods for estimating the difficulty of words, *bacteria* should be a much harder word for students to know. This is evidence of background knowledge and use driving student word knowledge. We then have a decision to make about the cultural reliance of reading comprehension tests for considering the *nature of* reading comprehension.

Two term that I will generally avoid are *definiendum* and *definien.* The definiendum is usually a term, symbol, or some other sign that is to be defined (Ogden & Richards, 1925). Alternatively, the definien is what we might call the definition, some set of words, images, or even attention directing action (e.g. pointing) that gives the definiendum reference and meaning. So, for instance, if I want to define *reading*, the term *reading* would be the definiendum and the definien could be something like “an activity that involves making meaning from a text” (though this is not a very specific definition, it gives the word meaning).

In the following subsections, I aim to more thoroughly introduce the case studies used in the remainder of this dissertation. I hope these subsections inspire thinking about how language in particular is important for each area of study.

3.2.1 Case Study: Definition of *Psychological Resilience*

In a review paper, Windle (2011) asks via the title of the paper, "What is resilience?" and notes that lack of consensus around a definition of resilience has both "methodological implications" as well as "strong implications for improving health and well-being" (Windle, 2011; p. 152) and sets out to solve definitional disagreement that is common in the field (also noted by den H artigh & Hill, 2022; Fleming & Ledogar, 2008; Luthar et al., 2000; Masten, 2001; Southwick et al., 2014b). Windle (2011) aims solve this definition of resilience problem through "concept analysis, literature review using systematic principles, and stakeholder consultation" (p. 153) and begins by providing linguistic and dictionary definitions of resilience. Windle eventually provides an analysis of how stakeholders use the word resilience including different definitions of the word (for instance, in the intro of this dissertation, we saw that authors use the word *resilience* to refer to very different phenomenon that occur at the individual and population level). The question is, what sorts of problem would Windle solve using these methods? I argue, there is no problem to solve, in fact, and we can use linguistic analyses to show this.

The sorts of analyses proposed and carried out by Windle are useful, but are solving a linguistic problem instead of, presumably, the one they hope to solve – identifying the core features to which the words resilience might refer. This seems circular, though – how would we know what *resilience* refers to without having a definition of it? It is likely, in the words of Wittgenstein, that resilience researchers have been “bewitched by their words.” Philosophy of language is not *just* about how language and words are used but also about the connection between historical account of words, even their etymology. Much of philosophy of language is devoted to what words refer to.

Windle (2011) and others (Ahern et al., 2006; Connor & Zhang, 2006; den Hartigh & Hill, 2022b) have devoted time to clarify a definition of resilience for the sake of measurement. Windle (2011) lands on a definition of resilience:

> Resilience is the process of effectively negotiating, adapting to, or managing significant sources of stress or trauma. Assets and resources within the individual, their life and environment facilitate this capacity for adaptation and ‘bouncing back’ in the face of adversity. Across the life course, the experience of resilience will vary.

This is not clearly a definition that would allow for the measurement of resilience – for instance, there are person level descriptors, multiple metaphors, and no clear identification of a unified concept let alone a property of persons. Windle has claimed that they have “clarified the nature of resilience.” I would argue, Windle has clarified the nature of how the word *resilience* is used in one Wittgensteinian language game. In effect, Windle has committed a use-referent conflation, confusing the question of “what is it?” with “how do people use the word *resilience*?”

Hibberd argues that, what she calls, scientific definition “is not a call to define words. It is not a lexical enterprise. At its most elemental, it is a call to answer the ontological question “what is it?” or “what kind of thing is it?” (Hibberd, 2019, p. 31). These are useful guidelines for understanding the goal of definition for the sake of research, and especially measurement. However, language is not irrelevant, either. Faccio, Centomo, and Mininni (2011) describe this nicely, invoking Wittgenstein and quoting Foucault:

> “the nature of mental phenomena is continuously reconstructed by the shapes we create in connecting representations to language: therefore, psychological objects have no stable quality or property, but rather “acquire truth from the methods and language devices that we apply in order to understand them” (citing, Foucault 1963, p.57)” (Faccio et al., 2011, p. 308).”

The argument is that mental phenomena are, to some extent, constructed by our linguistic practice. Thus, a fruitful endeavor for conceptual analysis for the sake of *scientific definition* is paying attention to how ideas, metaphors, and words may shape our everyday existences. This does not mean all words are equally useful in referencing something that is easy to research. For instance, Wittgenstein, as discussed in the previous chapter, was wary of language games. That is, he was worried that word meanings were not only fixed by what the words refer to but also how they were used in language games (Jost & Gustafson, 1998; Maraun, 2016; Wittgenstein, 2009). In the case of resilience, the language game is appealing to the metaphor of resilience (“bouncing back”, “process”, “navigation”) where the notion of the metaphor is taken from accounts of material resilience – the extent to which materials may bend but won’t break (this is a poor definition but nonetheless a paraphrase; Schiefer, 1933).

So what might be a fruitful way to consider resilience? Acknowledging the metaphor and the role of language will be a first step. In the case of resilience, for instance, its historical roots as a phenomenon of study in psychology are often traced back to Emmy Werner’s Kauai Longitudinal Study. Emmy Werner tracked approximately 700 children from

Kauai growing up in poverty and near poverty conditions for over 30 years. Approximately 2/3 of the children in Werner's study, according to the study, were exposed to particular risk factors prior to the age of two years old and developed behavioral problems, had delinquency problems, or early pregnancies. About a third of the sample were said to develop into "competent, confident, and caring adults" (Werner, 1995). This latter group of competent adults, Werner referred to as resilient. Werner connected the outcome of the resilient group to reports from that group's mothers that described their children as good natured, active, and easy to deal with. As an interesting point, resilience was not needing of a formal definition – it was doing loose work to refer to an outcome status. In elementary school these students were described by their teachers as having good problem-solving skills. Werner delves into protective factors that may have been prevalent among certain resilient youth. No explicit case is made for the assumption that resilience is a single, let alone composite, psychological property. The term resilience was a way to describe a group of people based on an outcome, not a particular within-person characteristic.

Resilience, linguistically, does not seem to have originated directly from the physical sciences – but one can see how the metaphor became problematic, attributing a potentially non-psychological attribute to individual within person claims. For instance, the authors of the CD-RISC, an instrument for "measuring" an individual's resilience define resilience as "embodying the personal qualities that enable one to thrive in the face of adversity" (Connor & Davidson, 2003). Another challenge facing the resilience realm, then, is the confusion between what is studied as a between person difference or within person difference. For instance, if resilience is simply about differences across people bouncing back from adversity (for instance, overcoming a challenge), this difference is simply a reflection of between-person differences. However, many have framed the so-called measurement of *resilience*, as in Connor & Davidson (2003) above, as a within-person conceptualization. Between subject differences can be the result of qualitative differences yielding a final quantitative difference (e.g., drawing on Borsboom et al., 2003 - had Danny had Andy's level of resilience, he would have resiled more. This casts a somewhat circular restatement about the state of an outcome instead of some causal connection between a person's property and their own outcome; this does not guarantee the same within person process).

The metaphor here, from bouncing back as a property of a person has been fully reified. Resilience might then be a dead metaphor which Chang (2016) described as metaphors "that are so ingrained in our way of talking that they are routinely mistaken as literal expression" (p. 111; also see Scheffler, 1963). The call here is then to focus on the phenomenon, instead of the word. In all, it is clear that language was part of the reasoning process about resilience but investigations focus on meaning of the terms instead of features of the world. Paying attention to this language will be important.

3.2.2 Case study: Student Reading Comprehension in NAEP

Contrast the resilience debates with debates around NAEP reading comprehension definitions and measurement of reading comprehension. In this case the debate has fewer unique specifications about the definition of reading comprehension, but, perhaps, just as much controversy. As mentioned in the introduction (1.2) of this dissertation, debates about what should be included in NAEP assessments to measure reading comprehension seem to combine issues related to the definition of reading comprehension and ways to measure it. It is at times hard to track the debate around NAEP because much of it has occurred outside of formal academic journal articles. On one side of the debate, educators and researchers hope to integrate research on sociocultural and cognitive factors in NAEP reading assessments. This includes, most controversially, universal design elements (UDE) proposed by the NAEP Reading Framework (*Framework)*. The idea is to provide scaffolds for students when certain words or certain ideas may pose challenges if student lack cultural experience or background knowledge. NAEP reading assessments will also include multi-modal or multimedia-based items and scaffolds that all together may have not been included in prior literacy or reading comprehension.

While NAEP is devoted to the notion of reading comprehension, the very idea of literacy keeps expanding, and it seems some of these new UDE based items reflect that. As mentioned previously, for NAEP updates, the governing board wanted to reflect updated research on:

> "(1) how social and cultural experiences shape learning and development; (2) how reading varies across disciplines; and (3) the increasing use of digital and multimodal texts" (National Assessment Governing Board (NAGB), 2021, p. 7).

Most controversially, the new framework for NAEP reading introduces what they call, "Information UDEs" which are texts that, in addition to the passage or test question, provide "brief topic previews" that provide necessary background information for understanding a text and "pop-up notes for definitions of obscure words or phrases that are not part of the comprehension target being tested" (NAGB, p. 9). Therein lies a heated source of debate that traces back to somewhat vague specifications about what is or is not included in the definition of reading comprehension that integrates *readers*, *texts*, *activities* (that readers engage in as they read), and *reader knowledge* (about the world, history, vocabulary, the particular text). The *Framework* says it has updated its broad framework in three ways to remain useful:

> "The first is how students' social and cultural experiences shape learning and development, including the learning and development of reading comprehension. The second is how reading varies across disciplines. The third regards the use of digital and multimodal texts." (NAGB, 2021, p.13).

Thus, NAEP defines reading comprehension:

"Reading comprehension is making meaning with text [bolding their own]**,** a complex process shaped by many factors, including readers' abilities to:

- Engage with texts in print and multimodal forms;
- Employ personal resources that include foundational reading skills, language, knowledge, and motivation; and
- Extract, construct, integrate, critique, and apply meaning in activities across a range of social and cultural contexts" (NAGB, 2021, p. 10).

Note, these are mixes of the definition of reading comprehension and, like the mistake made by Southwick in Southwick et al. (2014) about resilience, confuses causes of changes in reading comprehension with effects. For instance, the *Framework* states: "NAEP incorporates measurement of a wide range of factors that may influence reading comprehension" (NAGB, 2021, p.15) which raises the question about the debate of reading comprehension itself. Are we debating causes of reading comprehension or reading comprehension definitions? That is, reading comprehension is defined with "making meaning with text", meanwhile, most of the definition is about the process that *shapes* reading comprehension, implying that much of this definition is not about reading comprehension, but about causes of it. A definition of reading comprehension cannot include itself as a cause (Hume, 1740).

Additionally, like the confusion in the resilience realm, some of the change in reading comprehension makes it unclear whether we are considering between person or within person changes. For instance, NAGB (2021) cites Cronbach's worries that "individual psychological development" means "psychologists and educators would have to engage in systematic analysis of the interactions among the attributes of students and the characteristics of the settings in which their learning is fostered and assessed" (NAGB, 2021, p. 14). NAEP reading framework seems to be about between person claims, or, at least within group and not within person claims. In other words, there is an underlying, tacit orientation for using within person conceptualizations of reading comprehension, which, according to Borsboom et. al (2003), would hold between persons. Setting up tasks then requires thinking about how within person changes might be manifested.

Finally, NAGB (2021) incorporates multimodal texts in NAEP because of changing ideas about what counts as texts. So another challenge faced by NAEP's definitional work is the fact that the very thing that it intends to measure is changing. It is not just that we are learning more about the nature of reading or reading comprehension, it is that the very thing reading comprehension refers to is changing, perhaps due to cultural relevance (of course, if this is changing, it would not be the same property). So a question remains, should NAEP scores be compared from year to year? Here too, like resilience though not as transparent, NAEP reading may be bewitched by words. To expand further, in the case of NAEP, when the definition is changing, is the very thing intended to be measured changing (from previous generations of NAEP)? Answering this question may help answer some of the questions from the side of the NAEP definition debate that is against providing item scaffolding.

Unlike resilience, reading comprehension is still anchored in a relatively specific activity. Resilience is mostly anchored in scenarios being anchored by metaphorical representations of a phenomenon (facing a challenge of some sort). However, resilience is anchored historically by a particular study. NAEP's purpose for defining reading comprehension is rooted in cold-war-era aims.

NAEP was first administered in 1969 as a means to understand and compare student achievement across U.S. states in a standardized way (History and Innovation - What Is the Nation's Report Card NAEP, n.d.). Like the tests of Horace Mann's era described in the previous chapter, political motivations led to the development of NAEP. In particular NAEP came from Cold War era worries that the United States was not prepared to produce scientists or engineers to challenge the Soviet Union (Beaton et al., 2011). The 1950s version of the testing system that is now called NAEP was meant for use as school quality monitoring tools. NAEP was motivated by competition with the Soviet Union, and later in the early 1960s, as a product of school quality and equality of opportunity requirement in the Civil Rights Act of 1964, became an accountability tool. Many see NAEP as only gaining a stronghold in policy discussions following the publication of *A Nation at Risk* (National Commission on Excellence, 1983) which lamented the possibility of the United States' decline due to schools and educational systems more broadly not preparing students (a common theme that motivates testing programs[15]; Bourque, 2009). Since then, there have

[15] *A Nation at Risk* claimed*: "*America's position in the world may once have been reasonably secure with
only a few exceptionally well-trained men and women. It is no longer" (p. 10).

been numerous policy moves, including federal legislation signed by President Reagan in 1988 placing a governing board in charge of NAEP – primarily motivated by desires to compare state performance. Eventually, this led to legislating to set NAEP achievement levels in each subject area (Bourque, 2009).

Note, this history has not focused at all on *what* is being measured or how to consider measurement itself. Early attempts at standard setting involved selecting points on a range of scores estimated from student data and choosing anchor items that were around these points (based on student performance). Content experts described these anchor items relative to what they think these points on some scale correspond to in terms of reading development (Beaton et al., 2011; Bourque, 2009). In other words, the development of NAEP did not occur from a cognitive content perspective, at least not overtly. Like resilience, changing meaning and purposes of key terms are likely leading to debates that reify these terms – turning debates about content into debates about word meaning, though, less so in NAEP than in the resilience realm. One can also now see why a philosopher like Fricker (2007) might worry about power and transparency of this definitional effort in the *Framework*. Experiences of reading and comprehending text might be unarticulated in a reading comprehension definition that is relatively opaque like NAEP's (an example of hermeneutical injustice described above). Alternatively, without scaffolds, it is likely students cannot articulate their expertise or comprehension as *some* NAEP governing board members seem to fear.

The next several sections and following chapter on philosophy of language, science, and measurement are an attempt to frame these debates in a clear and transparent way. As mentioned in the introduction, a goal of this work should be *consensus* developing. I do not say the goal is *consensus*, because consensus should likely not be reached in highly uncertain areas. However, the work done should be such that we have a way to understand what we might develop consensus about. However, as will be discussed below, consensus and coherence cannot be the only criterion of good definition. Eran Tal (Tal, 2013) and Cartwright (2020) emphasize in measurement and in science (respectively) that coordination among researchers and scientists is how science operates. Tal asks a relevant question, speaking of measuring subjective well-being, but no less relevant to our examples in reading comprehension or resilience:

> “Attempts to validate questionnaires for measuring subjective well-being and quality of life raise something similar to the problem of coordination: are the questionnaires measuring what they should? Should the construct be defined in terms of the best-correlated questionnaires? It is doubtful whether these questions can be answered through a process of iterative stabilization similar to the one encountered in the standardization of physical quantities.” (Tal, 2013, p.1163)

The connection between measurement and language is interesting because both rely on reference and interpretation. For instance, in language, we will see that words in particular realms refer to specific things and can be used abstractly or not, refer to specific instantiations or universals (“author” referring to me or “author” referring to any given author with a given definition attached; on the measurement side, 3 inches can be a general idea that can refer to anything having that property of length or can refer to a specific thing being 3 inches long).

3.3 Cleaning it up: Increasing Transparency, Clarity, and Semantics in Property definition

As a reminder, the purpose of this chapter is about definition of measurable properties in psychology. The previous sections aimed to provide examples of problems that are linguistic and historical in nature in different stages of a research arc. A definition of measurement, or, at least, a characterization of what might be considered typical of measurement was provided in chapter 2. Definition for the sake of measurement is in part linguistic (after all, it is about the meaning of words), and part empirical – about whether what those words refer to have some mode of existence that have real-life instantiations. The next place to move toward is to pay more attention to how we can classify the different ways language is working. .

3.3.1 Introduction to philosophy of language for measurement: sense and reference

Our goal here is guided by Hibberd:

> “Although we use words to propose or state what kind of thing we think it is, a definition references the what-it-is-to-be that kind of thing — its principal features or structure — in order to delimit it from other kinds and to make possible a systematic study of it and its connections” (Hibberd, 2019, p. 31).

When we have words like “resilience” setting the direction of research, it is not always clear what the word refers to or delimits. The notion of delimiting or referencing, comes from Frege (1892/1948) who was concerned with, in the linguistic realm, logical bases of words and their meanings. He introduces “Sense and Reference” with a challenging scenario of identity. That is, when we make a statement of the form `a = a`, this is so obvious as to be uninteresting. If we say `a = b` then perhaps we say something interesting but it might be also obvious. For instance, when we say resilience is not the same thing as thriving (Carver, 2010) this is interesting because it means that the word *thriving* refers to a different phenomenon than the word *resilience*. In the realm of latent variable modeling, it is often debated what a factor "is", or what a factor in a latent variable model refers to. This is, in effect, a question of reference. The problem of an identity, when a = a or when the "morning star = evening star"[16] or that grit and conscientiousness are the same (e.g. Ponnock et al., 2020), what is being said is that the referent of two different terms are the same. That is, the term, or the sign, "grit" refers to the same thing as the terms, or the sign, "conscientiousness." This equality is interesting because it is making a claim about advancement in knowledge.

In science, it is an accomplishment to realize that what we once thought were two different entities areare in fact the same entity. This requires both theoretic and model-based accounts of the thing being observed converging. Hibberd's statement above is then essentially arguing that for the sake of science, words, parameters in a model, or something similar should refer to something. That something is, in effect, what the definition should describe. Though, Frege was thinking of proper names, certainly, this refers to the names of objects or idealizations like "the meter."[17] When we use the word "moon" says Frege, we "presuppose a referent" an and are not referring to the speaker's individual conception of the moon. This presupposition allows us to ask, "what is resilience?" and try to solve this question via empirical means. The only way for us to answer the question, "what is resilience" is by already knowing to some extent, what resilience might be and we have to, in

[16] The cat is also on the mat.

[17] It is not necessarily settled that idealized terms in measurement models refer to anything real, though, it is taken in this paper that which are referred to in measurement, properties and their values, are real and referred to.

effect, point to it. However, resilience has no immediately obvious referent given the various definitions above.

Why we have a shared notion of what the moon is, or why terms like *resilience* or *thriving* or *reading comprehension* have references in speech, is a sense of the term, effectively, their meaning. For instance, there may be "excess" meaning beyond the referent. The sense of morning star is effectively Venus in the morning. The sense of evening star would be seeing Venus in the evening. This may seem unimportant but clearly the sense of a word can also confuse us, allowing us to attribute the same term to different things or two different terms to the same thing. Meaning is not a function of properties or objects, then, it is a function of words and intersubjective understanding.

Frege claimed this when he said that the sense "is grasped by everybody who is sufficiently familiar with the language or totality of designations to which it belongs" (p. 210) and a given referent could have multiple senses. This is important for latent variable modeling and the notion of "constructs" and to what they refer. For instance, Kane (2008) says, "Observable variables are defined in terms of domains of possible observations and therefore have little excess meaning" (p. 105). Depending on what Kane means by variable (e.g. Markus, 2008), this is most likely incoherent. Variables, if we are to interpret them as what terms refer to (e.g., a synonym for property or attribute here), have no meaning without us attributing meaning. In that sense, a property cannot have a meaning until we name it. For instance, we say that reading comprehension is a property of a person and we ascribe the predicate or property, *property of a person* to the phrase *reading comprehension.* Further, the term *reading comprehension* should refer to something. Kane, perhaps, has a non-realist perspective about measurement, but since, earlier Kane (2008) mentions that his coffee cup is plainly observable in a way that something else is not, admits some forms of realism. Of course, language in validity theory has a history of not worrying whether “constructs” refer to anything. The logical positivist roots of psychometrics going back to “construct validity” and Cronbach and Meehl (1955) permits constructs to have, effectively, no referent and only meaning (e.g. Lovasz & Slaney, 2013; Slaney & Racine, 2013).

Cronbach & Meehl said constructs got their meaning from their place in a nomological net, where definition was about meaning – “setting forth the laws in which it occurs” (Cronbach & Meehl, 1955, p. 290) and must be either a so-called observable property or be related to something in a nomological net where at least some of the properties are observable. This is related to meaning, because a nomological net, at best, is a model or representation of system. And like other sign systems, (e.g. Ogden & Richards, 1925), one can see that this does not require a referent. A classic semiotic triangle based on Ogden and Richards is presented in figure 3.2 below. Cronbach and Meehl (1955), are only concerned with the left side of the triangle connecting thoughts and ideas to terms or symbols. There is no requirement that the term or symbol is connected to a referent. This requires that a construct is only embedded, in some capacity, among observables.

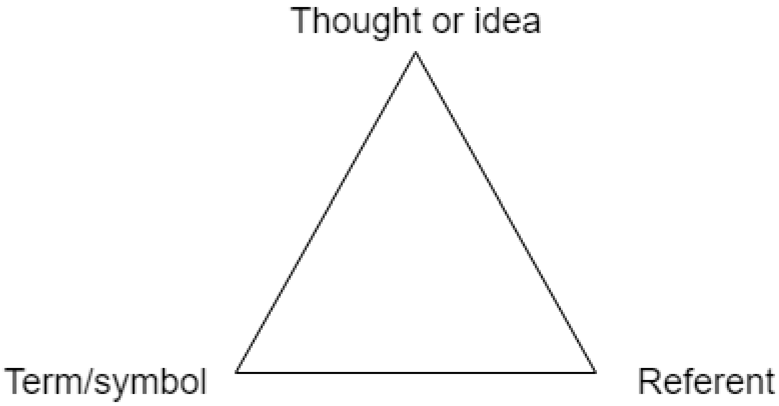


FIGURE 3.2 *THE SEMIOTIC TRIANGLE FROM OGDEN AND RICHARDS. A THOUGHT OR IDEA MIGHT BE CALLED A CONCEPT IN FREGE*

To some extent, we see this line of thinking in Frege where he refers to the referent as an “object perceivable by the senses” (p. 212), though, perhaps, perceivable by the senses could be expanded to admit perception via scientific tooling (see section on pragmatic realism, below).

3.3.2 Solving measurement problems with language

Using the language of Giordani and Mari (2012) in any given setting involving measurement, one is trying to solve a “measurement problem” and a measuring instrument acts “as a selector, interacting with the object under measurement with respect to a given quantity, the measurand” (Giordani and Mari, 2012, p. 2146). In this conception, as noted in Chapter 1, measurement involves an interaction between that which we would like to measure and the instrument. In physical sciences, this might involve bringing in physical laws that describe how a measurand changes an instrument indication as well as other properties that might be involved in the interaction of a measurement tool with an object's property (e.g. Grégis, 2015; Tal, 2019).

Here, for ease, properties that we are interested in are considered universals. Arp et al. (2015) define universals as "entities in reality that are responsible for the structure, order, and regularity — the similarities — that are to be found there. To talk of universals is to talk of what all members of a natural class or natural kind such as a cell, or organism , or lipid , or heart have in common" (p. 14). Properties, whether of persons or of hearts or of rooms and spaces, are not immutable (necessarily). For instance, my height can change. In other words, a universal is something that two things can have in common. I can have the same height as someone else, but I also generally have *height* as a property of myself. This implies, to some extent, that "the complex bio–psycho-social systems of interest to psychologists exist or occur independently of their observing, thinking, talking or writing about them" (Hibberd, 2019, p. 31). Though, certainly, Hacking's looping effect might render this in partial truth. The connection between continuous psychometric testing and reporting on skills that are named renders those skills important to the community at large, gives teachers and parents something to call those skills and act in an "evidenced-based" way, perhaps in turn causing the creation of more tests of these skills. This is an intersubjective creation.

Returning to sense and reference, we can see that a referent is not guaranteed to a word even if it has multiple senses. Alternatively, one word can refer to many things or two words can refer to the same thing. In essence, Frege's scenario is in figure 3.3 below. (and others not mentioned here such as Russell; see Kripke for another version of this, though, with some critique). In these diagrams, squares are terms that refer to the circles, stand ins for the referents. In the top left panel A of figure 3.3, different terms are used to refer to the same entity. In top right panel B, we have two individual terms referring to different entities (here, “tc” and “rc” were used to note that these are not the words, but some portion of reality). In the bottom panel c) one term refers to two different things. A speaker might use the term not knowing they were referring to two different things (which seems to be the case in resilience research), or, as mentioned in chapter 1, the context in which the speaker uses the word might dictate the referent of the word. This is an example of pragmatics in language use, to some extent. We cannot require a perfectly descriptive all-encompassing language for communication (and it is probably not possible) because communication would likely take forever. However, for scientific purposes, we cannot rely on general understanding else we run into the problem of resilience researchers.

Finally, in the bottom right, is an interesting scenario. Odds are, many of us have the conception of a unicorn in our head. We can picture a horse-like thing with a single horn and maybe the horse-like thing is even white. Saying that, a unicorn may not have a physical instantiate to go along with it (Quine, 1948, uses the example of Pegasus). However, we might say that unicorns or leprechauns have an existence in fairy tale or myth. For now, however, we will treat this as an empty reference. The word refers to something that cannot be realized in any particular way outside of an idea. This perhaps points out that starting with a discussion of “what is real” when doing research may not be so productive.

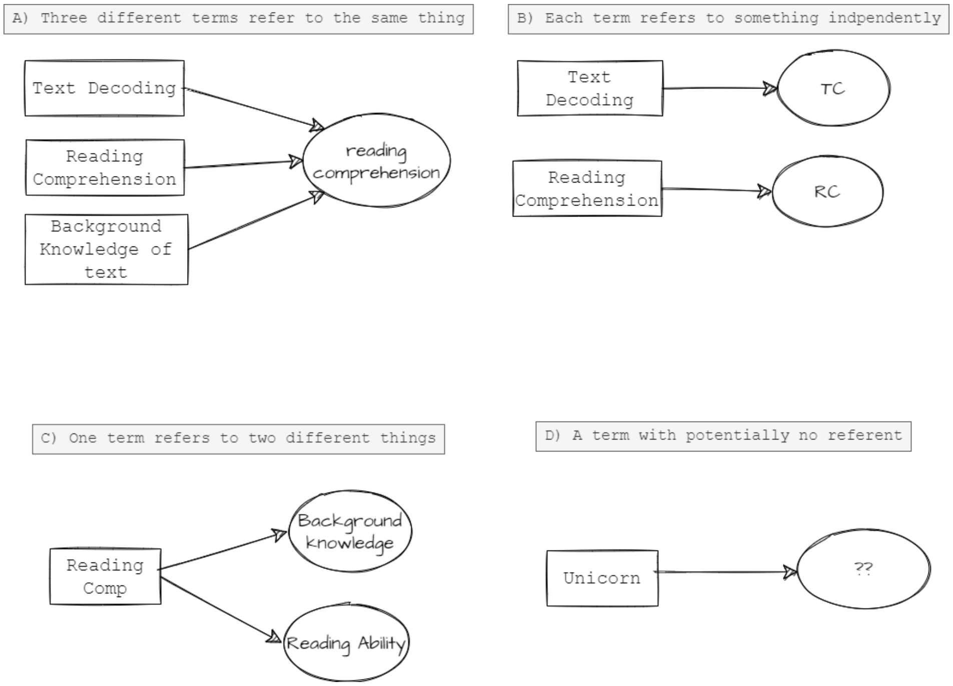

A) Three different terms refer to the same thing
Text Decoding
Reading Comprehension
Background Knowledge of text
reading comprehension
B) Each term refers to something indpendently
Text Decoding
TC
Reading Comprehension
RC
C) One term refers to two different things
Reading Comp
Background knowledge
Reading Ability
D) A term with potentially no referent
Unicorn
??

FIGURE 3.3 *THESE FIGURES EXPRESS DIFFERENT WAYS WORDS MIGHT REFER TO REFERENTS. IN THIS VERSION, SQUARES ARE WORDS OR TERMS AND CIRCLES ARE MEANT TO BE THE REFERENT THEMSELVES (OF COURSE, STILL BEING REPRESENTED BY WORDS AND DIAGRAMS).*

Note everything above may be reliant on speaker intentions. Authors such as Kripke have noted that there are scenarios where the speaker may use words (or names) to refer to, say, one person even when that is not the name of that person. Not represented in the figure is the scenario where we get the sign and the signifier confused. In this case, we might call this concept-entity conflation or even, use-referent or use-mention conflation mentioned in the previous chapter. We might confuse a noun or phrase for the thing itself (e.g. “resilience is x because that is how people use the word resilience” or “resilience has three e’s and is the capacity to bounce back from a challenge; Lovasz & Slaney, 2013).Wittgenstein might generally refer to this as being bewitched by our words (Jost & Gustafson, 1998; Wittgenstein, 2009, 2010).

3.3.2 Types of Definition – What makes a good definition?

Hibberd was quoted above as saying that a good definition states what it is to be that thing. In other words, a word will be connected to what it refers to by its definition or ascription. Further, in the words of Hibberd (2019), for definition about entities for study, there are some things that definitions do not do that are of interest in this dissertation. For instance, they are not stipulative or nominal. This means that word meanings are not just randomly assigned to words but have some purpose for being assigned to that word. In another light, we are not just concerned with the meaning of the word but describing the phenomenon itself. One can imagine, however, that to some extent, all definitions are stipulative, though, using a word out of context of a language game might confuse. The broad swath of this definitional effort is lexical. Further, the goal is not to come up with examples of the thing, though this may be helpful or to point and define. It seems interesting that at times, stipulation is of interest and other times it is not.

Stipulative definitions for research are necessary to the extent that words need some assignment of meaning (that may come from use; Scheffler, 1963). However, for research, the goal is not to determine the meaning of a word. The goal is to determine or discover universals, enabling consensus building because there are few criteria aside from linguistic rules to determine when a word is properly defined. For instance, it would be odd to say that reading comprehension is a golden retriever. However, when debating whether resilience *is* bouncing back from adversity or an internal psychological process, there is no answer because both things can exist in the world and we are just discussing how to assign word meaning which has no real resolution because of different language game rules. This becomes clearer if we replace *resilience* with a variable name or made-up word. Arguing whether *x* refers to bouncing back from adversity or some psychological process is not a question about the phenomenon.

What we are instead interested in is demarcating the process or phenomenon. The NAEP reading comprehension debate exemplifies where there are a mix of stipulative definitional concerns and phenomena-based concerns. One side of the NAEP debate argues that reading comprehension, as a phenomenon, is what Kintsch and van Dijk (1978) describe. This description of reading is about the cognitive process that goes into connecting a text's parts (e.g. placing words together, or even morphemes, into sentences and ideas) and then into a coherent idea about what the text is trying to say in its entirety. Kintsch and van Dijk (1978) formalize this as micro processing and macro processing. In the micro context, they theorize that readers piece together hierarchies of rule following about reader deletion of somewhat irrelevant propositions in the text, reader generalization of series of specific propositions to the more general one, and construction of global, logical, or well accepted facts from a series of propositions make up the text processing model.

In other words, there is a micro level and a macro level, and they relate to each other. In this view, reading comprehension refers to these phenomena and background knowledge is external to the phenomenon of reading comprehension, but is, nonetheless important to the macro processing. It is, in effect, what is processed. As mentioned above, another side of the reading debate questions the nature of reading as "sociocultural." For instance, Steiner & Bauerlein (2020), in an online journal publication, argue that this sociocultural perspective desired in the 2025 (now 2026) *Reading Framework* over emphasize sociocultural context. They argue that the "hows and whys and strengths and weaknesses of reading comprehension change [given the framework] from one sociocultural context to another, leaving some students in potentially "unfair" starting points, depending on the relationship between their backgrounds and the texts in question" (Steiner & Bauerlein, 2020). Ignoring the voracity of the argument for a moment, it is clear, given the purpose of measurement in general, that both sides of this debate worry about the effect of admitting or not certain elements into the world of reading comprehension and how this will lead to problematic, untrustworthy measurement. We can see that this leads to some stipulation (of the term *reading comprehension)* still, since meaning needs to be assigned to the term, but this might be what some call a precising definition – it is limiting or expanding what can be included in a relatively well accepted realm. Though, there is still an ethical element amounting to normative claims and measurement aims.

For now, I aim to point out how a stipulative definition puts a researcher in the position of defining a word in the way they see fit. This is important at times, when a term with many senses needs to correspond to an actual referent in the world. Of course, the danger is demarcating a non-universal, arbitrary class of entities. There is also a matter of power. For instance, by including background knowledge as part of the demarcation of reading comprehension as a property, one might worry that this definition will privilege groups with particular knowledge. Answering the result of the question invoking the counterfactual - *had students without that knowledge been provided necessary knowledge* - might provide some insights into what is of interest for investigation.

To be clear, it is the position here that aims of measurement are to work with properties or attributes that are non-arbitrary or universals. Defining the term, "Denverite" as all people who happen to live in Denver, ride bikes, and are in graduate school is an example of a stipulative definition that identifies something arbitrary and is stipulative. In the case where I misuse the term *reading comprehension* or *Denverite*, people might be confused because there is a typical use of these terms. If I instead use the words in their typical way and use the phrase, "A person who lives in Denver is a Denverite" – this would be a type of *reportive* definition (e.g. Scheffler, 1963).

For measurement, we do not want to be *only* arguing about stipulation or reportative definitions. Instead, it is more fruitful to pay heed to the phenomenon or set of observations that might be common occurrences. Ogden and Richards (1925) describe the first step for settling on definition as finding a common starting point – a common realm. If you start with *resilience*, note its history and the change in meaning of the term – it now has many more senses beyond Werner's study. The goal for arguing about what the phenomenon that the term *resilience* refers to is to facilitate a common frame of reference. In essence, if you want to start with *resilience* as a human psychological process, stay in this realm. Moving to Werner's realm, "bouncing back" – is a different one. Of course, Ogden and Richards (1925) also note that "whenever a term is thus taken outside the universe of discourse for which it has been defined, it becomes a metaphor, and may be in need of a fresh definition" (p. 111). So, even in reading comprehension ability, when using terms like "apply" or "integrate", we have to make sure we are still in the same discourse or frame of reference.

3.3.3 Avoiding a Bewitching: Paying Attention to Metaphors and Language in Use

As mentioned before, the key to measurement coordination and consensus building starts with iteration and improvement of the terms, models, and background assumptions. Understanding measurement results requires that we collectively agree on the focus of measurement. In the words of Lakoff and Johnson (1980), metaphor is "understanding and experience one kind of thing in terms of another" (p. 5). The problems we run into come about when we stick to preferred definitions for reasons unrelated to the property of interest. For instance, if I say resilience is a process, that likely implies that that is what I'm interested in studying. However, because the word *resilience* has so many senses, many seem to think they aim to study the same phenomenon. For measurement, this is especially problematic because it is about a specific property and the value of the property. *Resilience*, as mentioned above, is a particularly good case of metaphor. The *sense* the term gains from the metaphor, something spatial or something extensional, implies a particular structure (something that goes up and down). This dictates how we conceptualize something – it has looking at something in a particular way.

Metaphor is quite common in educational and psychological research. For instance, *depression* is a spatial metaphor. Cognitive load is also something of a spatial metaphor (activities take up space *in* the brain and have a *weight*). Metaphors are not inherently problematic, though– Lakoff and Johnson (2008) note that we cannot communicate without metaphor. Brown (2003) traces the productive use of metaphor in science, starting with the atom and how, at one point, atoms were thought of like specs of dust. These metaphors inspired experimentation based on the mental model. However, the metaphors were iterated on, improved, concretized, and abstracted. Brown (2003) warns us that metaphors "also pose a danger: attachment to a particular model [posed as a metaphor] can inhibit thinking in other, possible more productive ways about the system that is being studied" (p. 25).

Wittgenstein (1953/2010) additionally warns us that the problem of clarity in definition, is "solved, rather, by looking into the workings of our language, and in such a way make us recognize these workings … the problems are solved not by giving new information, but by arranging what we have always known" in order to "battle against the bewitchment of our intelligence by means of language" (Wittgenstein, 2010, sec. 109). Said another way, studying language in use might indeed help us figure out how we are confusing ourselves when we are thinking about how scientists use that language. To exemplify this, let us consider *resilience.* It was argued above that investigating the meaning of *resilience* is non-empirical (unless one is a linguist), and investigating the definition of resilience is investigating something we already know since it is defined in the dictionary or has a given set of rules in use (e.g. a language game mentioned in chapter 2). According to Wittgenstein, when you ask a question like "what is resilience?", provide many definitions that are all plausible and non-mutually exclusive, you are playing a language game:

> "there is: The tendency to look for something in common to all the entities to which we subsume under a general term…Games form a family the members of which have family likeness. Some of them have the same nose, others the same eyebrows, and others again the same way of walking; and these likenesses overlap" (Wittgenstein, 2009, p. 106).

Language games help us communicate, as do metaphors (which might be a special case of a language game), but they get in the way when we play them using different rules across communities for the same term. Even though *resilience* is the same term, it has definitions expressed by different researchers that vary wildly (Southwick et al., 2014; Windle, 2011). We can ask the question "what is *resilience?"* likely because it presupposes resilience already exists when in fact this is relying on its metaphorical extension to another realm. But from its instantiation, the concept of resilience was a matter of metaphor starting with Werner's study and the meaning of the term in that context were people who "succeeded" despite adversity – it was a somewhat arbitrary label invoking the notion of bouncing back.

In that sense, the term resilience, as applied to some psychological process is an empty reference where we can use the term because of reference to a common structure – orientation and spatial metaphors from negative experience (down) to bouncing back (up). This leads to measurement instruments that involve positing resilience as something continuous in some form, without any question about the nature of the structure of resilience (e.g. Heilemann et al., 2003; Smith et al., 2008; Windle et al., 2011; Xie et al., 2016)

What about the realm of reading comprehension in NAEP? We see in the *Framework's,* definition that reading comprehension is a process in which the student might:

> "Extract, construct, integrate, critique, and apply meaning in activities across a range of social and cultural contexts" (NAGB, 2021, p. 10).

Two obvious metaphors in this are extracting or constructing. (ignoring others such as, "integrating", "in activities", "across a range"). It is posited that "in the mind", students are "extracting" (some sort of physical or spatial metaphor from the text, implying mining or moving information) and constructing (some sort of metaphor involving building a structure) and doing so in a certain context. It is not clear how this now relates to what is intended to be measured in NAEP reading since these are many different processes. However, to contrast with resilience, the metaphors are not the core of the definition of reading comprehension or the core of the debate. In the reading comprehension realm, there is a particular activity (reading) that is anchoring the debate. However, the word "comprehension" and metaphors above as being part of comprehension may be a form of bewitching. That is, because comprehension in reading comprehension is somewhat ill-defined (or even the term *literacy*), it is not too specific for any context so it can fit in many places. In this way, we might be debating across each other where one side of the literacy debate is arguing in one framework that is focusing on the cognitive side of comprehension and is not wrong *per se*, and the other side is arguing in an older or different framework or use of the word comprehension and literacy.

Notice, the idea of sense and reference stance introduced here is inherently realist – there has to be something “out there” to which we are referring (though, not necessarily). However, how do we demarcate a phenomenon when it is socially constructed or if it is socially constructed, does that mean it is “real” and measurable?

3.4 Cleaning it up: Considering Measurement, Realism, and Pragmatic Realism

While not an explicit definition of measurement, Tal (2019, p. 870) says measurement consists of “the coherent and consistent attribution of a value region to a parameter in an idealized model of a process, based on the final states (`indications`) of that process” and hence involves “intervening and representing.” This model-based view, which is adopted in this dissertation, means that some simplifications must be made – measurement is not an epistemic tool without a model of what we are measuring. Tal (2019) argues that measurement is an intervention because it requires a design process for constructing something that will measure what one intends to measure while also requiring the recording and interpretation of measurement outcomes (e.g. – instrument indications). To measure, one requires a model of measurement “constructed from the theoretical (or pre-theoretical) and statistical assumptions” (Tal, 2019, p.870) to make sense of the indications of measurement instruments. For instance, to make sense of a student’s reading ability as estimated from a reading test, a model of how the student responds – a response process – would be required. To the extent that there is doubt about the model’s correspondence to the way the item responses on the reading test are generated, there will be doubt about using those scores to make claims about students – the indication of the instrument will not be trusted. For instance, the trustworthiness of the way a weight scale interacts with an object's weight comes from background assumptions and a model of the process of weighing. If this process is somehow in doubt, then a lot of uncertainty will be attributed to the measurement claim.

An example of this model-based account comes from Sherry (2011) who is responding to a challenge posed by Michell (e.g. 1999, 2008). Michell critiques educational and psychological measurement for not having "experimental tests known to be specifically sensitive to the hypothesized additive structure of the attribute studied" (199, p. 216). Michell thinks measurement requires showing that, if one were to make (continuous) measurement claims about something (temperature, reading ability), one requires showing that these attributes are indeed quantitative in *nature* first. Sherry (2011) retells the tale of the temperature measurement in which early researchers, notably Joseph Black, made progress via "*treating* [italics in original] temperature as a continuous quantity" (2011, p. 517) instead of first testing whether temperature is a continuous quantity. Black used this assumption to "construct upon new thermal concepts, useful for explaining and predicting thermal phenomena" (Sherry, 2011, p. 517). Sherry argues that temperature as a quantitative property fit nicely into thermodynamic laws which in turn explained or predicted other phenomena.

According to Sherry and Chang (2004), Gay-Lussac was able to claim that volume of gasses at a given pressure was proportional to temperature – meaning that readings of thermometers relying on the expansion of gases as a form of temperature indications could be interpreted (between 0° and 100°) and measured non-arbitrarily. Additionally, this finding could be used to generate the ideal gas law which "provided the basis for the thermodynamic temperature scale, today's standard." (Sherry, 2011, p. 517). One can see that the trustworthiness of thermometers was reliant on a model of temperature and its relationship to expanding gas – specifically, it was reliant on the ideal gas law. In that sense, the trustworthiness and understanding came from a model of interaction between thermometers and temperature. While there was not an accurate account of the nature of temperature (or heat) – this was the start of an account of temperature – and there was certainly a recognition that an account of the cause of different temperatures was required (Sherry, 2011).

While Sherry (2011) credits this account of the measurement of temperature to a philosophically pragmatic attitude, there was also a certain correspondence to reality (some might call this "pragmatic realism" from e.g. Guyon et al., 2018; Maul, 2013). Chang (2004) and Tal (2019) attribute the accomplishment of measuring temperature to a view called coherentism. Roughly, the idea of coherentism in the philosophy of science is that a theory is considered true (or accepted) when it fits into either an already existing network of ideas or makes sense in terms of acceptable phenomena. A problem with coherentism as pointed out by a Chang (2007) is that "any internally consistent system of knowledge is equally justified" (Chang, 2007, p.4). Instead, Chang (2007 – and to some extent, also in Chang, 2004) proposes an alternative that he calls progressive coherentism. The idea merges some foundationalist doctrines with coherentist doctrines. Namely, it is foundationalist because progressive coherentism says scientific progress is made by accepting some "system of knowledge without ultimate justification" but coherentist and empiricist in the sense that it attempts to improve that system of knowledge by exposing it to thought, experimentation, and other "lines of inquiry which can … refine and correct the initially affirmed system" (p. 5). Science, posits Chang, advances via "epistemic iteration" (2007, p.18) in which knowledge is improved upon iteratively – with previous knowledge used in new stages of investigation and improved upon given epistemic aims. This position, however, does not require a devotion to scientific realism. In fact, the emphasis on epistemic aims might involve things that do not require a scientific realist account. In other words, it is not necessarily the case that coherence is building towards a truth – Chang says as much – but it is nonetheless an attractive description for the innerworkings of science. This position also has political implications – the direction of research will be built upon accepted knowledge bases – including some voices and excluding others (e.g.Kuhn, 1970). Nonetheless, Chang's progressive coherentist approach offers useful direction. It is contended here that coherentist theories are useful for understanding, perhaps historically, why or how scientists justified their beliefs that is not purely falsificationist (Popper, 2005) .

Thagard (2007; citing Goldman, 1999) alternatively gives an account of truth that is something like a correspondence theory: "a representation such as a proposition is true if and only if it purports to describe reality and its content fits reality" (p. 29). However, coherence theories state that the important relations are between mental representations as opposed to correspondence with external reality (Thagard, 2007). Meanwhile, Chang seems skeptical that truth is something that is even worth working towards. This seems fair given what some have cast as "pessimistic induction" (Newton-Smith, 2002, p.14). The pessimistic induction idea is that since most (or all) theories in science have historically been, at some point, accepted to be false (for instance – Newton's laws being replaced by relativity, caloric theories of heat having explanatory power only to be replaced), why should we accept a theory as true? Chang's epistemic iteration from coherence theories of truth may provide a counter to the pessimistic induction objection to scientific realism, ironically. In doing so, this links epistemic iteration instead to a potential description of a *methodology* for a certain form of scientific realism where theories are ever improved upon in a march toward the truth.

Devitt (1997, 2005) attributes to scientific realists the commitment to (presently) unobservable entities being responsible for observed phenomena and scientists seek to find or use these phenomena for explanation. Scientific theories link these unobservable entities to observed phenomena and these phenomena exist mind independently. Thus, the pessimistic objections argument states that since science has posited past phenomena that did not exist, it is likely that current theories with posited phenomena have problems. However, as Devitt argues, citing Lange (2002), one problem is that, given the way science operates, false theories might “turnover much more quickly than true ones” (Devitt, 2005, p. 265) so a given surviving theory may be more likely to be true. Devitt posits, what he says, is a stronger argument:

> “Scientific progress is, to a large degree, a matter of improving scientific methodologies often based on new technologies that provide new instruments for investigating the world. If this is so . . . then we should expect an examination of the historical details to show improvement over time in our success ratio for observables” (Devitt, 2005, p. 265).

In essence, Devitt is proposing a defense of realism based on a form epistemic iteration. Even if past posited unobservable phenomena turned out to be false, iteration of knowledge improves the likelihood of truth. In the words of Chang (2004) – an aim of epistemic iteration is self-correction – meaning there will be plenty more past theories than present accepted theories. Chang’s primary objection is to truth as justificatory criteria for posited theories (or unobserved entities) – and in general, moves away from a correspondence theory of truth.

It is therefore suggested that Borsboom and Hibberd's realism do not necessarily conflict with Chang's *progressive coherentism* or epistemic iteration if epistemic iteration is viewed in consideration of what scientists do with realist intent. It is noted in all cases that scientists perform some prior necessary conceptual or theoretical analyses, even if not realized. Borsboom (2005, 2008a) and Borsboom, Mellenbergh, and Van Heerden (2004) consider the justification of claims for having measured in scientifically realist terms. For instance, in their definition of validity, Borsboom et al., (2004, p. 1061) say a test is valid for measuring an attribute if the "attribute exists" and "variations in the attribute *causally* produce variations in the outcomes of the measurement procedure." Further, they argue that truth of the "ontological claim [for instance, the status of what is measured as existing and in what form it exists] is logically prior to the process of measurement itself" (p. 1062), since measurement involves measuring *something* believed to exist. It would be incoherent to say one has measured something that does not exist since measurement involves the interaction of a measuring instrument and an attribute (Mari et al., 2012b; Maul et al., 2018b).

Chang's (2016) own recent form of realism, something he calls pragmatic realism – a position not that different from Putnam's (1987; Putnam called his version, initially, *internal realism* and later, pragmatic realism) – de-emphasizes correspondence theories of truth and emphasizes taking actions with certain aims but exposing them to our current understanding of reality (not dissimilar to Devitt's improved instrumentation above). It is realist in the sense that there is still an emphasis on truth in the form of entities existing. Chang (2016) says, "A statement is true in a given circumstance if (belief in) it is (necessarily) involved in a coherent epistemic activity" (p. 115). Chang also makes reference to a reality – "a putative entity should be real if it is employed in a coherent epistemic activity that relies on its existence and its basic properties" (Chang, 2016, p. 116). This definition also draws on Hacking's entity realism and trust in the senses as a foundation for understanding what it may mean for us to admit something is real. In other words, how might we admit some entities but not others (e.g., psychological attributes of people)? Chang does not desire an *anything goes* mentality. One answer, provided by Hacking (1999), is that certain social constructions can be considered real, say to a person, if it makes a difference in their lives. He uses the example of the classification of woman refugees. It is a social construction because a person is only a woman refugee given various social norms about gender, countries, borders, and the legal classification of someone as a refugee. However, it does not have to be this way. Thus, classification, or property of a person as a refugee is real because it changes how that person lives – even if they do not know the term "woman refugee" – so that property interacts with the world in a certain way. Hacking calls this an *interactive kind*. Social constructionism according to Hacking (1999) is:

> "…Various sociological, historical, and philosophical projects that aim at displaying or analyzing actual, historically situated, social interactions or causal routes that led to, or were involved in, the coming into being or establishing of some present entity or fact" (p. 48).

Pragmatic realism and Hacking view of social constructionism complement each other in that both are still, to some extent, trying to make some connection to reality, just sometimes, we create that reality as people. In that sense, one could deal with Borsboom's definition of validity in a pragmatic realist sense – the justification for existence of an entity would then just require a specific line of reasoning. If anything, this provides some credence to Lord & Novick's claim that (ignoring some of their terminology) something can be assumed to be measured when observed results correspond to the target of measurement. Only, in this case, it is beyond mere assumption as Lord & Novick emphasize.

The pragmatic realist position is attractive, and there are clear cases where pragmatism and realism seem incontrovertible (e.g. using tools for successful navigation would make one hard pressed to argue against the realness of certain measured entities that allowed for successful navigation). In the social realm, dogma is always a risk. One could argue that measured IQ is real because one can take certain actions based on those IQ "measurements". Perhaps a counter would be in terms of IQ's coherence – breaking down the idea of IQ making it incoherent, since people with different skills can accomplish different things. Alternatively, one could argue that IQ's realness is incontrovertible given correlational evidence. Perhaps this is the case, but the idea of social constructionism's unmasking role (Hacking, 1999) may be of use. The idea may be made false or untrue or similar via removal of the concept in some way such that different criteria are used to discuss intelligence. Saying that, this perhaps shows that early investigation into a phenomenon should not necessarily start with the concept of "realness" but, instead, posit strong use cases where differences in measurement results could lead to actual different actions taken. In the words of Chang (2016, p. 119):

> "When Hacking says that positrons are real, or when I say phlogiston is real, the sense of it is that a specific part or aspect of that unspecified overall Reality is somehow being captured in our conception. And this parsing-out of Reality is crucial in any kind of cognitive activity. If we cannot identify sensible parts (or aspects) of

> nature, we cannot say anything intelligible, make any kind of analysis, or engage with nature in any specific and directed way. So we have no choice but to worry about whether we are able to do the parsing well."

Perhaps an objection to this is that the conceptions of those things as "real" may in fact be their connections to what replaced those concepts as science improved. In fact, one might not be able to do much with the concept of phlogiston now without invoking its replacement property. As usual, perhaps this is not so much an expression of a philosophy of science in general as a statement about how science progresses, a meta-methodology. We admit into our realities that which we can move forward with.

It is hard to imagine trusting measurements, research, or other forms of analysis by which we make public policy decisions without some foundational bed rock related to the truth of the matter or the way the world is. Even implicit in psychometrics, regardless of philosophical orientation is a notion for realism. For instance, testing agencies are not just concerned with prediction of future e outcomes or a prediction task else testing companies might create some maximally predictive index combining test scores, grades, parental education, and parent income, among other things.

3.4.1 Underdetermination and realism: all we have are models

Another important consideration is the Duhem-Quine thesis or the underdetermination of theory by evidence idea. The basic premise is that evidence in research is always amenable to alternative, if not nearly infinite interpretations. In this sense, we cannot base our beliefs only on the data that we have (see, for instance, Stanford, 2021). There is a technical version in the structural equation modeling literature related to fit – where perfect fit of model to data can still come from a mis-specified model (for instance, Hayduk, 2014). We see this in statistics where the same correlation can be found when the structure of relationships between two variables are very different. In that sense, all we have is our beliefs about the world to narrow down plausible models or reasons. We can reject models, research findings, or data all together when it conflicts with the way we know the world even if evidence in that moment says otherwise.

In this sense, we will model our realness and workflows with this picture of science in mind. Epistemic iteration, coherence, logical analysis, and observation will be interwoven, working toward these ideals. We think about measurement from a model relative position that has posited entities interacting with measurement instruments. These posited entities should be considered real but with some skepticism. We must justify these entities that feature in a measurement process and how we justify them will be through an iterative, sometimes circular, process. Unlike Chang, we will not ditch a correspondence theory of truth, that good models or theories somehow correspond to reality and that understanding reality, operating in the real world is a necessary goal as well, though, we agree with Chang in the sense that there indeed may be other goals along the way. A powerful tool for considering whether to admit some property into our ontology is whether we can, in any way, meaningfully intervene on that property, for instance, in a classroom or via social means.

3.4.2 Alternatives to essentialist accounts of properties and views of science

The views above also assume some sort of essentialist account of properties in that a given kind will always have the same set of characteristic. Hibberd (2019) introduces the concept of the homeostatic property cluster (HPC) as a non-essentialist account of definition in which certain properties (Hibberd uses the term, “kinds”) have some core features and some of these features appear in some contexts but not others. Hibberd argues that this may be popular because it accommodates the notion of changing concepts or kinds and vague boundaries (or properties). Hibberd argues that the HPC account entails that “some kinds are defined by a cluster of features that regularly but not exceptionless co-occurrence; and a set of factors (causal homeostatic mechanisms) that maintain their systematic co-instantiation or clustering, factors that provide some necessary cohesiveness or stability to the cluster” (Hibberd, 2019, p. 40). For instance, we may see that reading comprehension requires one form of background knowledge in one reading setting but not another or that something like reading comprehension from 50 years ago is the same as reading comprehension now even when it involves integrated video in a text as is common in online newspaper articles. Hibberd rejects the HPC conception because if a kind changes it is no longer the same kind as before. Alternatively, one can shift conceptions as not just temporal change but contextual change. Consider the reading comprehension debate in NAEP about incomparable scores across time. The HPC account renders these scores comparable if the set of properties involved in reading changes. Ultimately, this is confusing and I tend to agree with Hibberd, though I cannot obviously rule out HPC accounts, and mereological positions generally.

Like Hibberd, I think the HPC account may reflect an epistemic stage that requires iteration to clarify the boundaries of a property. Unlike Hibberd, I think the HPC account may be important in some settings but not others. For example, it does seem like reading or literacy is a changing concept. However, a purely cognitive account might admit that micro and macro processing (Kintsch and van Dijk, 1978) would not rule out current conceptions of literacy or reading comprehension that involve a new integration of video into text, for instance, as being a form of “reading comprehension” necessarily. In this sense, integrating video and text for a single student reading was simply mostly unrealizable before personal computers. This alone cannot rule out the “ability to watch videos” clustering with “text decoding”, “vocabulary knowledge”, and knowledge of semantics and pragmatics. As noted, earlier, though, definitional uncertainty may be a form of epistemic uncertainty. Hibberd argues that HPCs more likely reflect that HPC is an accommodation of “early to mid-stages of scientific definition” (Hibberd, 2019, p.49). I argue that this may be a permanent state of scientific definition (in the words of Hibberd), and we are instead relying on forms of epistemic iteration – this may also provide a rebuttal to the notion that we must show that a psychological attribute is quantitative before proceeding with measurement.

3.5 The Problem with Latent Variables

The lengthy section above was needed for justifying the treatment of what we define as real and measurable and the referents of our words, especially in the psychological sphere, because, as noted in chapter 1, definition is often cast in an operational light. However, it is common in educational or psychological measurement to refer to or invoke *latent variables* as those things that are measured or to say that a latent variable **is** a factor, construct, hypothetical entity, unobservable, or something similar. However, the present unobservable status of an entity is not a defining feature. A general question may be asked – “what are latent variables?” But this question is incoherent without context.

Borsboom (2008b; Borsboom et al., 2003) notes that unobservability is a statement about our present epistemic access. For instance, what one person in the same time and place may treat as unobservable, another person may treat as observable (for instance, through one person having a microscope and another person not). Maxwell (2009; first published in

1962) noted, effectively, that what was once considered *unobserved* (or theoretical)[18] may become observed through improved instrumentation. The uncertainty around that observation may be what induces the notion of unobservability. The implication being that, perhaps, treating something as permanently unobservable is an unsafe assumption.

Further, the notion of a latent variable is only meaningful in statistical or mathematical analyses. In other words, latent variables cannot be what are measured. This was noted by Maraun & Halpin (2008) when they said that a variable "is simply a rule/map/function… and, hence, cannot coherently be placed on the dimension of observable to unobservable" (p. 114). To provide more detail, a random variable is defined in intro probability texts as:

> "Given an experiment with sample space *S*, a *random variable* (r.v) is a function from the sample space *S*, to the real numbers $\mathbb{R}$." (Blitzstein & Hwang, 2014, p. 104)

This effectively means that the mapping to numbers is not necessarily a constituent of reality (Maraun & Gabriel, 2013). Though, while this may be an overly strict interpretation of the term *variable* (Markus, 2008), in agreement with Maraun & Halpin (2007, p. 9), I cannot take Borsboom's (2008, p. 9) definition that an observed variable simply means that the value of a variable as it is realized can be "inferred with certainty from the data." For instance, in the case of a continuous value, the data may only be stored to a certain number of digits – but as a continuous value, and hence a member of an uncountable set, may not have a

[18] It's important to note who Maxwell was writing to and when. He was countering the logical positivists, a group of philosophers working to make, among other things, a theoretical and observational distinction such that reference to unobservable or, alternatively, theoretical entities was somewhat meaningless. This distinction is interesting in the realms of psychological or educational measurement where unobservable entities are posited but sometimes those entities might be understood in terms of operationalism.

realization as a rational number – hence its value can never be known without some uncertainty, even if miniscule. Instead, it is maintained here, drawing on the discussion of epistemic iteration and coherence above, we treat something as observed when, the number of background assumptions that support the scientific work to observe something are few and there is intersubjective agreement about those assumptions. Hence, something well accepted, could be modeled as latent to handle rounding or measurement error.

Consider a scenario where we need to measure the distance between a point on a bicycle wheel hub to another point on the same hub for determining the spoke length necessary for building the wheel. It would be a costly mistake timewise to get the measurement wrong. So, I take several measurements and each time get a slightly different estimate. I may use the internet to see values given for similar (or even the same) hub. Using all this information, I may take an average of my measurement results. Suddenly, I have constructed a latent variable model depicted in figure 3.4 below. Note, estimation of the mean can be construed as a latent variable model, but it seems odd to do this. However, there is a “true value” that is unobserved while there are observed measurement results/observed data perturbed by measurement error.

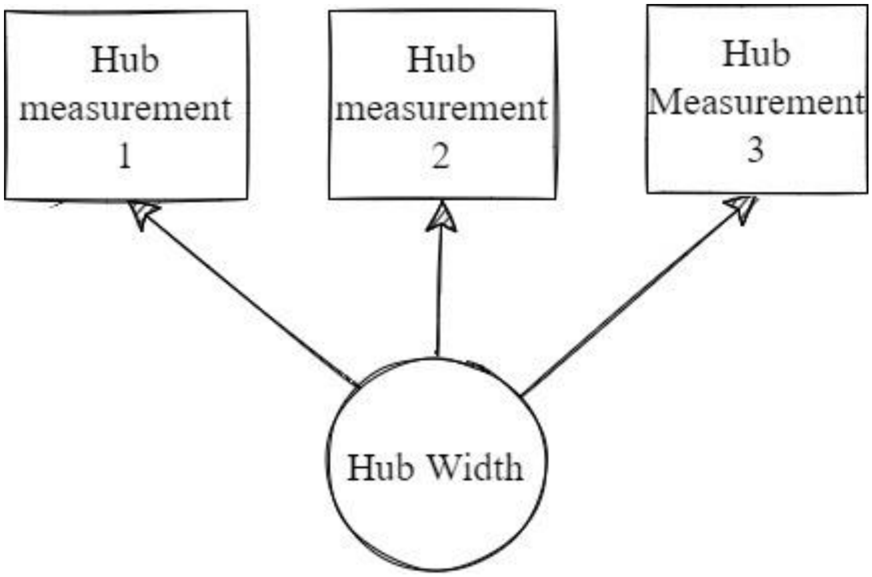


FIGURE 3.4. *A LATENT VARIABLE CONCEPTION OF A WIDTH MEASUREMENT*.

One can now invoke hub width as a latent variable. Our/my collective confidence in measurement results are not certain but they are high.

To bring back Chang (2004), he says "observability is an achievement" (p. 87). Part of that achievement requires picking out what is to be measured or working around consensus building. I hope the sections below help with consensus building.

3.6 Recommendation: Specify your Measurand for Transparency and Fairness

One prevailing problem with the measurement in NAEP reading comprehension debates and resilience research is the lack of a model of the measurand from any side of the debate. In the case of resilience, I hope I have been able to show that the debate is primarily linguistic instead of substantive whereas the NAEP debates are more substantive but influenced by different human values. Paying attention to sense and reference may be a useful place to start for resolving both problems. Let us first, then, try to carve out a referent, ideally, the measurand, and map how it might be related to the student reading performance.

In the case of NAEP reading comprehension debates, the term *reading comprehension* should have a single referent. The boundaries of those terms may be fuzzy, but as mentioned above, this is both natural and a good thing. Alternatively, it may be the case that we cannot articulate a single property, and that is also fine. Drawing on Kintsch and van Dijk (1978), background knowledge matters but we can consider it an influence quantity in NAEP reading. Unfortunately, it is unclear whether this is a quantity as in the GUM that is thought to change the key property itself or if it is just the measuring instrument indication. We might draw a causal diagram of the theoretical modeling process to help us understand the focal measurand based on language. For instance, Steiner & Bauerlein (2020) wrote that:

> "One of the most powerful and consistent findings about reading comprehension is that it depends on the background knowledge that a student possesses. . . Yet the whole point of the NAEP reading assessment is to tell educators how well states and school districts are teaching students to read the language they will encounter in the real world. That report, in turn, tells us how well prepared American high school seniors are for college and the workplace. If NAEP adopts the sociocultural model of reading, the scores for certain student populations will almost certainly improve, but the scores themselves will lack any predictive value" (Online, no page number).

There are a lot of assumptions in this quote, but we can gain an understanding of what the term "reading comprehension" (though, not explicit in this sentence) might mean to those arguing against scaffolding in NAEP reading. It seems reading comprehension is something along the lines of older NAEP reading framework definition that involve "understanding written text" and "Using meaning as appropriate to type of text, purpose, and situation" (they also note that it is a dynamic cognitive process; Steiner & Bauerlein, 2020). Steiner &

Bauerlein (2020) also argue that this definitional language is commonsensical despite the fact these terms are mostly undefined. This appeal to common sense should be problematic, given the discussion about being bewitched by words and metaphor. For instance, the combination of background knowledge and cognitive processing here creates something of a combination of properties or, it could be two different properties. What then makes this combination of cognitive processing and background knowledge *reading comprehension* instead of something like, “history” or “literature knowledge?” We might draw the relation between the property or properties of interest and the resulting reading performances (scored via items) as in Figure 3.5, below.

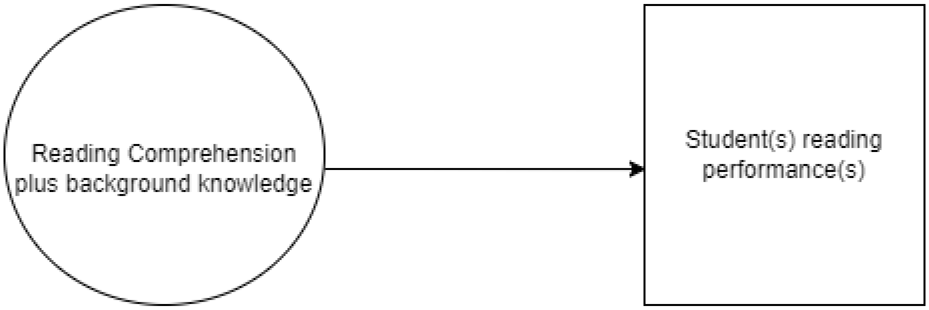

Reading Comprehension plus background knowledge
Student(s) reading performance(s)

FIGURE 3.5 STUDENT BACKGROUND KNOWLEDGE VERSION OF READING COMPREHENSION PROPERTY. THE ARROW IS NON-DETERMINISTIC, BUT A STRAIGHT LINE IS USED TO NOTE THAT THERE IS NOTHING NECESSARY TO BE POSITED IN BETWEEN.

Figure 3.5 shows that this version of reading comprehension is something of a black box (circle?) interpretation of reading comprehension ability – whatever it is that causes a reading performance is reading comprehension or a classic latent variable formulation. This does not answer how background knowledge or cognitive processing combine.

Unfortunately, in the NAEP case, in the *Framework’s* definition of reading comprehension, there are numerous sub targets deriving from the components of reading comprehension. Ironically, in this mode, there is no explicit discussion of the relation of these sub disciplines to each other or to the primary measurand (reading comprehension ability). In most cases, they are concerned with how to handle background knowledge in terms of item creation which, while tacitly implying what the definition or what may or may not be in the definition of reading comprehension, is not an explicit definition. Those parts that go into reading comprehension (presumably; though are these causal?) are listed as comprehension targets: “locate and recall”, “integrate and interpret”, “analyze and evaluate”, “use and apply” (NAGB, 2021, 17). These are situated in contexts of: discipline (e.g. literature, science, social studies); a given reading purpose (developing understanding, solving a problem, and more specific purposes); and types of texts and text features. The question is whether each comprehension target is some form of reading comprehension, a part of reading comprehension in a mereological or even homeostatic property cluster sense, or a formative sense (e.g., these are measured parts that formatively demarcate reading comprehension).

Alternatively, is reading comprehension to be measured at the Kintsch and van Dijk (1978) macro level or micro-level? Is each change of context of a passage also a new property? Here, the referent of “reading comprehension” seems to be overrun by the sense. However, this aspect from NAEP is far more specific than the black box account. Drawing out the relation between the measurand and resultant responses, we see in Figure 3.6, how we might conceptualize this in the formative sense described above. In figure 3.6, there is a dashed line connecting background knowledge to reading comprehension ability while a solid line connecting it to the student reading performance indication. The dashed line to the reading comprehension ability circle aims to show that it may be the case that background knowledge is not just an influence quantity that influences the indication but may affect the measurand as well. In the top left, we have a cascading/tree-like structure that dictates that each reading occurrence will be in each context. If we are to imagine reading comprehension as a function or disposition (Arp et al., 2015), then reading comprehension can be considered the same property in each setting – reading comprehension is a function involved in these contexts (e.g. $f(context)$ where f is reading comprehension or, discourse comprehension Kintsch, 1988). I am by no means committed to this, this just seems plausible.

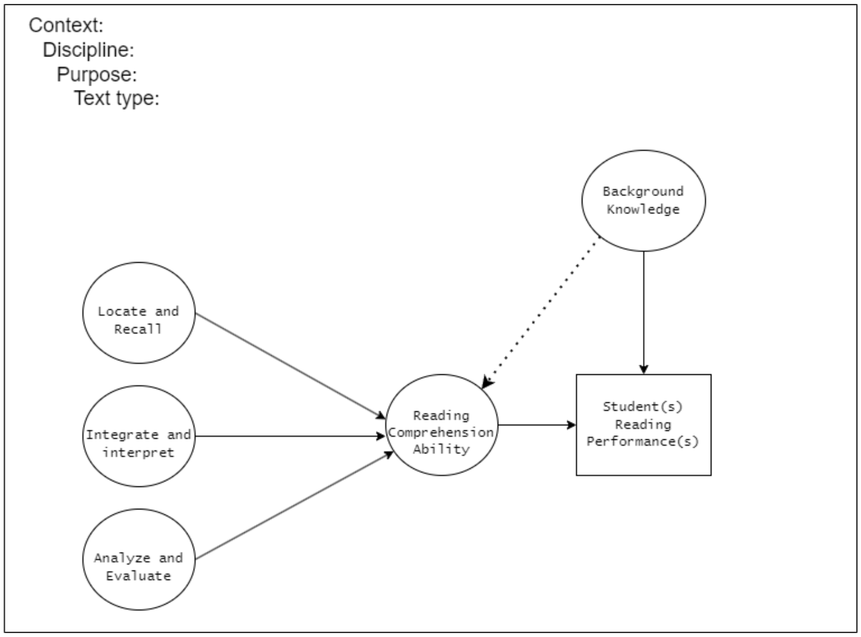


FIGURE 3.6 *DEPICTING THE RELATIONSHIP BETWEEN THE MEASURAND (READING COMPREHENSION) AND READING PERFORMANCES AS DESCRIBED IN THE 2026 NAEP READING FRAMEWORK.*|

While this may identify the structure or what we want to measure, we still somewhat lack a definition that enables us to measure. The definition can occur in many ways including textual. One way would be to follow the lead of Kintsch and van Dijk (1978) or even Kintsch (1988), but a definition of reading comprehension then would have to decide whether one wants to measure reading at the macro level or micro level. Clarification could take place in several ways, some of which are more stipulative and others are more use-based forms of definitions. If we are to borrow from Kinsch (1988), we might have a somewhat strange (because of my own attempt to keep this short and lack of subject matter expertise) definition that reading comprehension = def. the extent to which a reader integrates ideas within a text together as opposed to constructing disparate ideas only in each sentence (or something like that). This only is part of the comprehension model and requires the definition of the term *integrate*, which is clearly a metaphor. We might notice, though, we still have a choice to make about what is evidence *of* integration.

The above distinction directs us to the next point that at times it is hard to define simply in terms of a few sentences. One way to help our definitional work would be to use a construct map (Wilson, 2005), though, perhaps we could call it a property map. A construct map is rather simple. Adapted from Wilson, it might look something like Figure 3.7. The person side of the construct map would include descriptions of people who can do things at varying levels of the property. For instance, "can connect arguments in chapter 1 to arguments in chapter 2" (this is likely too vague but might work with our integration example 2. Observations would be examples of instantiations of doing this. This contributes to transparency and can be a tool that is provided and shared with stakeholders and allow for consensus building. It also encourages multiple communities' involvement in constructing or defining properties, allowing for iteration. Empirically, we can combine the construct map with the DAGs such as figures 3.6 and 3.5 above to better understand the observational ordering along the continuum. We might re-order our passages based on this, but also consider how background knowledge and use-experience influence even simple vocabulary. One might question then, is there any form of vocabulary that is not some form of background knowledge? Here, we have empirical and conceptual elements interacting. The next chapter in the dissertation will be an attempt to provide another example of how using statistical modeling, thinking from metrology, and philosophy may be combined.

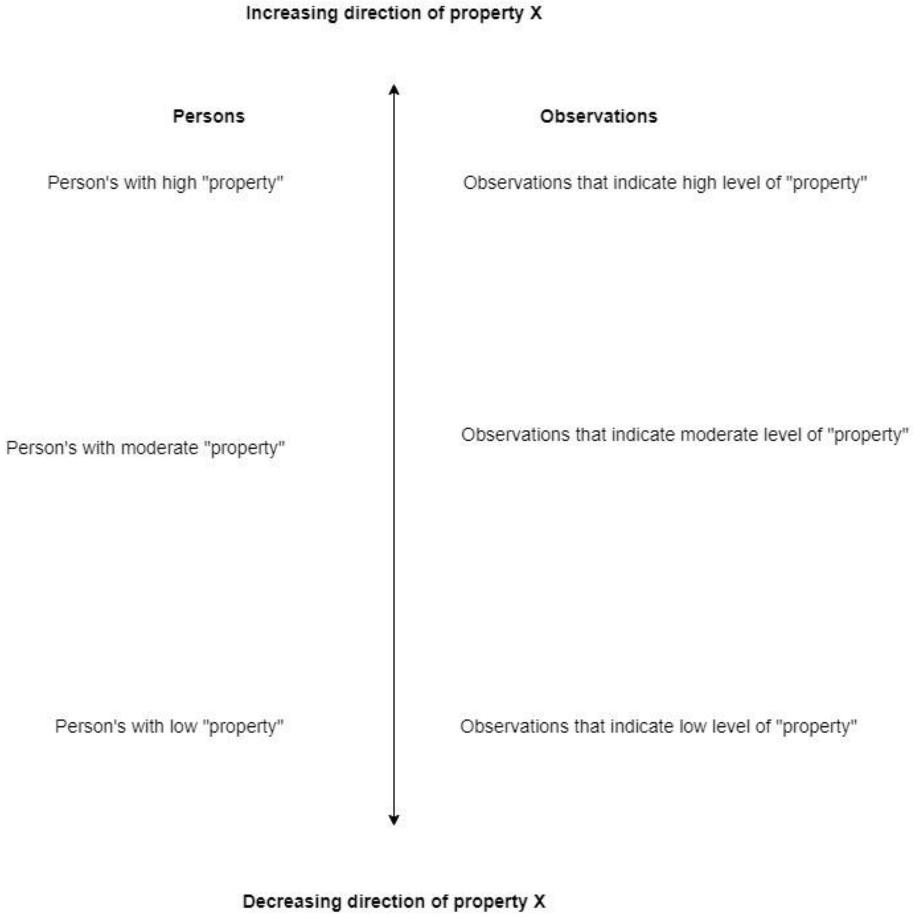

Increasing direction of property X
Persons
Observations
Person's with high "property"
Observations that indicate high level of "property"
Person's with moderate "property"
Observations that indicate moderate level of "property"
Person's with low "property"
Observations that indicate low level of "property"
Decreasing direction of property X

FIGURE 3.7 *EXAMPLE OF A CONSTRUCT MAP ADAPTED FROM WILSON (2005).*

However, there are two different specification of what the construct map implies and one overarching question for the property in general. One important matter is to not assume that the property you are working with is either continuous or even ordered. It is possible that the construct map may lead you astray. This is something that may be testable but will also change a definition of the property of interest. Secondly, we should be careful to specify whether we believe something like the construct map holds only between people or represents some sort of developmental pathway. It is possible for instance, that while the continuum or ordering described above may describe differences among people at differing levels of ability, the way in which students move through these levels themselves may in fact be different from this construct map. This is something that likely requires a fair amount of iteration, research, and conceptual work. Both question hint at an even larger problem. There is no guarantee that the property exists as we think it does. This would imply that the description or what we know, definition, or very nature of the property lends itself to not being measurable.

This lack of measurability is realized in the resilience definitional debates. I hope it is plain to see that there are so many varieties of definitions in resilience research, construct mapping may not be of much use to decide on the correct definition of resilience because they have many different referents. In the outcome realm, resilience is defined by a life outcome such as no longer being in a particular adverse position you were earlier. This is sort of a between person view that, if anything, is a classification task. Alternatively, there are the internal, psychological positions, which are psychological and defined as an attitude, belief, or propensity to do something. Defining the measurand here is not about picking the right definition to match to a word. Instead, one should identify the area of research that they are working in and try to identify the primary property while ignoring the loaded word *resilience*. Each definition of resilience above are simply identifying potentially different phenomena that need articulation.

Finally, with construct mapping, it is important to avoid circularity. A challenge with mapping attitudinal properties or things we have typically relied on self-report tools for measuring, is that observations on the construct map may often be quite similar to person property descriptions (on the left side above). This is a problem as circularity is a problem in other forms of definition. Including the word one is defining in the definition of itself does not provide clarity. The property of interest, in other words, is not the observations by which we make inferences about the property of interest.

3.7 Discussion, Limitations, and Paths Forward

This chapter has been an attempt to articulate a philosophy of language to discuss, transparently, the properties we intend to measure in the human sciences for measurement. I hope it is clear how the language we use is important for not only how we see the world but also how we can be fooled by our language. In this way, measurement, like language, has a semantic and syntactic structure that relies on pragmatics (or use). Here, I hope to summarize some final recommendations based on what is written above from the perspective of the methodologist or statistician and the subject matter expert working together. For both groups, it is important to understand and identify the referent of a term. Definition in science is not just about what a word means but what in the world we are trying to study. If a term has multiple referents, it is likely you have to be more specific than colloquial usage allows. While we cannot communicate without metaphor, it is also important to pay attention to it. It can confuse us so that we all believe we are saying the same thing. Remember that we are human, and science and measurement are human activities. As such, always express uncertainty about both what we are trying to measure and the results from a measurement activity. Here, to some extent, following Slaney and Garcia (2015), remember that measurement results are merely estimates at best of a particular property value (more on this in the following chapter). We should not reify and should explicitly say that “our estimates of student abilities from the text comprehension instrument and the text decoding instrument have a correlation of…”. If you are a methodologist, it is likely quite important to talk about that what we are measuring are properties and not latent variables in order to avoid confusing the terms in the model with properties in the world. For the methodologist, it is important to remind the substantive researcher that statistical tools do not define but they can help define. Data itself is constructed, even if unarticulated, by a model of how to instantiate that data such that it is evidence of something. For the substantive researcher, it is worth reminding the methodologist that the inferences from models are not only the job of the substantive researcher – in order to construct and make inferences about a model, understanding of the model is important. Some level of understanding on the part of the methodologist is necessary – invoking again, the idea that models are built for a purpose.

Pay attention to the history of a term and how it came to be. This history gives a clue as to the nature of the discourse and shifting meaning as well as how researchers might mix meanings with older uses of the term and newer uses. Consider whether these new uses shift the referent of the term. As methodologists or statisticians, do not define from data derived from surveys or tests. This likely ignores some of that discourse, and, given uncertainty, may not warrant any definition. That is, the possibility of data as being indicative of some underlying phenomenon requires background knowledge not found just in the data set. Otherwise, if you do name factors, taxa, classes, or similar, note that these are stipulative definitions – assigned names to statistical artifacts, and are *not* themselves, the phenomena of interest.

Importantly, do not talk about properties, constructs, latent variables, or dimensions or similar having a “meaning.” This is a form of concept-entity or use-referent conflation. By saying something means something, we are confusing what we are claiming about what the property we would like to measure is and leads to a complete lack of clarity about what the property is. Finally, we should pay attention to how we use our words in multiple ways. Even if we can have a conversation with a word like *resilience,* we can mean completely different things by it. In that way, we need to see when a word is being used to refer to two different things because there is some similarity in how it is used. We cannot build any form of consensus or agreement this way. If you are not sure what something is, remember that this is different than a latent variable. In this case, we may just not know enough about some regularly observed phenomenon to know that it is measurable. Related to this notion is the idea that we cannot define via statistical analyses. Definition is external to analyses which are model and value laden.

There are major limitations to the views expressed above and in this chapter. There are some undefined terms, or terms that describe a lot more attention – not the least of which is the term *truth*. Above, for the sake of clarity (I hope), we have times mixed the aims of science as an institution or cultural phenomenon (without defining) and the aims of the individual scientist. In part, I portrayed the aims of science as grounded in a certain form of realism that holds the truth in high regard, but clearly, a certain “constructive empiricism” might be still be admitted for the individual scientist that admits theories without being strongly committed to the entities posited in those theories as existing or being true. This adheres well to Chang’s views, that scientists, in the form of constructivist empiricism perhaps, might have goals about empirical adequacy first (e.g. results from research *seem* to fit well with posited theories that are specific enough to have a syntax and semantics; (van Fraassen & others, 1980; c.f. Monton & Mohler, 2021)This is a large topic, but it has been mostly glossed over. For instance, we may object to the notion that measurement results are of a property and refer to anything real or true. This would require a discussion about why universals are real or what models are true.

Additionally, there is a tacit assumption, though a dangerous one, that we can understand each other's experiences and words in all scenarios. I do not think this is true. In this case, it is worth questioning whether something that requires understanding of vague terminology or specific feelings in order to measure a property yields trustworthy measurement results. That is, it is not clear that psychological properties that are thought to be attitudinal or emotional are in fact possibly felt in the same way across people. We see the problem of epistemic injustice raise its head here – the very articulation of a particular feeling or sensation or experience may simply not exist for some people. This may shoehorn certain experiences into a term or phrase that is not quite appropriate. In other words, think carefully about the ways language may be used differently with very different meanings.

Alas, one of the bigger conundrums is simply what counts as a referent in psychology or education. In section 3.4, I tried to piece together how we might admit psychological attributes into our ontologies. In this way, we can measure them potentially – which requires they be universals. However, this explanation of psychological or educational attributes only partially refers to anything substantial. In some sense, the referent of psychological attributes are behaviors that seem to have some regularity. On the other hand, this relies heavily on the observations and not the causes. Though things like physical instantiations need not be the only referent (e.g. biological or neurological-based referents), at some point, connecting these referents back to the human body seems necessary.

This chapter has not provided an empirical basis for differentiating properties or quantities. For instance, how would we know that text comprehension or text decoding are different from macro-processes of reading comprehension? Tal (2019) calls this the practice of individuating quantities. The following chapters will be an attempt to provide methods for reasoning via statistical results about the properties or descriptors that may or may not be part of the larger property of interest. We will repurpose models for detecting differential item functioning (DIF) in order to detect and abduce what influence quantities might be in a given testing setting. This is a special case of epistemic iteration. The GUM, clinically, acknowledges that "the evaluation of uncertainty is neither a routine task nor a purely mathematical one; it depends on detailed knowledge of the nature of the measurand and of the measurement. The quality and utility of the uncertainty quoted for the result of a measurement therefore ultimately depend on the understanding, critical analysis, and integrity of those who contribute to the assignment of its value" (Joint Committee for Guides in Metrology (JCGM), 2008, 3.4.8, p. 8). I have to wonder to what extent the necessary background knowledge about measurands and what skills or professional senses exist in educational and psychological measurement to express uncertainty. If we cannot express what we do not know, how can we express what we do know? Clearly, the psychometrician alone cannot be a source of that skill or knowledge. I hope this chapter and the following help bring about some notions of building the transparency that the GUM itself calls for.

References

Ahern, N. R., Kiehl, E. M., lou Sole, M., & Byers, J. (2006). A Review of Instruments Measuring Resilience. *Issues in Comprehensive Pediatric Nursing*, *29*(2), 103–125. https://doi.org/10.1080/01460860600677643

Arp, R., Smith, B., & Spear, A. D. (2015). *Building ontologies with basic formal ontology*. Mit Press.

Arya, D. J., Hiebert, E. H., & Pearson, P. D. (2011). The effects of syntactic and lexical complexity on the comprehension of elementary science texts. *International Electronic Journal of Elementary Education*.

Beaton, A. E., Rogers, A. M., Gonzalez, E., Hanley, M. B., Kolstad, A., Rust, K. F., Sikali, E., Stokes, L., & Jia, Y. (2011). *The NAEP Primer (NCES 2011-463).*

Blitzstein, J. K., & Hwang, J. (2014). *Introduction to Probability* (Vol. 112). Chapman and Hall/CRC.

Bollen, K. a. (2002). Latent Variables in Psychology and the Social Sciences. *Annu Rev Psychology*. https://doi.org/10.1146/annurev.psych.53.100901.135239

Borsboom, D. (2005). Measuring the Mind. In *Measuring the mind: Conceptual issues in contemporary psychometrics*. https://doi.org/10.1017/CBO9780511490026

Borsboom, D. (2008). Latent Variable Theory. *Measurement*, *6*, 25–53. https://doi.org/10.1080/15366360802035497

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The Theoretical Status of Latent Variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bourque, M. (2009). *A History of NAEP Achievement Levels: Issues, Implementation, and Impact 1989-2009*. https://eric.ed.gov/?id=ED509389

Brown, T. L. (2003). *Making truth : metaphor in science*. University of Illinois Press.

Cartwright, N. (2020). Why trust science? Reliability, particularity and the tangle of science. *Proceedings of the Aristotelian Society*, *120*(3), 237–252. https://doi.org/10.1093/arisoc/aoaa015

Carver, C. S. (2010). Resilience and Thriving: Issues, Models, and Linkages. *Journal of Social Issues*, *54*(2), 245–266. https://doi.org/10.1111/j.1540-4560.1998.tb01217.x

Chang, H. (2004). Inventing Temperature. In *Science And Technology*. https://doi.org/10.1093/0195171276.001.0001

Chang, H. (2007). Scientific Progress: Beyond Foundationalism and Coherentism. *Royal Institute of Philosophy Supplement*, *61*, 1–20. https://doi.org/10.1017/S1358246107000124

Chang, H. (2016). Pragmatic realism. *Revista de Humanidades de Valpara\'\iso*, *8*, 107–122.

Cicchetti, D., & Rogosch, F. A. (1996). Equifinality and multifinality in developmental psychopathology. *Development and Psychopathology*, *8*(4), 597–600. https://doi.org/10.1017/S0954579400007318

Connor, K. M., & Davidson, J. R. T. (2003). Development of a new resilience scale: The Connor-Davidson Resilience Scale (CD-RISC). *Depression and Anxiety*, *18*(2), 76–82. https://doi.org/10.1002/da.10113

Connor, K. M., & Zhang, W. (2006). Resilience: Determinants, Measurement, and Treatment Responsiveness. *CNS Spectrums*, *11*(S12), 5–12. https://doi.org/10.1017/S1092852900025797

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

den Hartigh, R. J. R., & Hill, Y. (2022a). Conceptualizing and measuring psychological resilience: What can we learn from physics? *New Ideas in Psychology*, *66*, 100934. https://doi.org/10.1016/J.NEWIDEAPSYCH.2022.100934

den Hartigh, R. J. R., & Hill, Y. (2022b). Conceptualizing and measuring psychological resilience: What can we learn from physics? *New Ideas in Psychology*, *66*, 100934. https://doi.org/10.1016/J.NEWIDEAPSYCH.2022.100934

Devitt, M. (1997). *Realism and truth*. Princeton University Press.

Devitt, M. (2005). Scientific realism. In *The Oxford handbook of contemporary philosophy*.

Earvolino-Ramirez, M. (n.d.). Resilience: A Concept Analysis. *Nursing Forum*, *42*(2). Retrieved June 7, 2017, from http://www.nursingacademy.com/uploads/6/4/8/8/6488931/resilienceaconceptanalysis.pdf

EngArc - L - Modulus of Resilience. (n.d.). Retrieved June 14, 2017, from http://www.engineeringarchives.com/les_mom_modulusofresilience.html

Faccio, E., Centomo, C., & Mininni, G. (2011). "Measuring up to Measure" Dysmorphophobia as a Language Game. *Integrative Psychological and Behavioral Science*, *45*(3), 304–324. https://doi.org/10.1007/s12124-011-9179-2

Fleming, J., & Ledogar, R. J. (2008). Resilience, an Evolving Concept: A Review of Literature Relevant to Aboriginal Research. *Pimatisiwin*, *6*(2), 7–23. http://www.ncbi.nlm.nih.gov/pubmed/20963184

Frege, G. (1948). Sense and Reference. *The Philosophical Review*, *57*(3), 209. https://doi.org/10.2307/2181485

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

Giere, R. N. (2009). An agent-based conception of models and scientific representation. *Synthese 2009 172:2*, *172*(2), 269–281. https://doi.org/10.1007/S11229-009-9506-Z

Giordani, A., & Mari, L. (2012). Measurement, models, and uncertainty. *IEEE Transactions on Instrumentation and Measurement*, *61*(8), 2144–2152.

Goldman, A. I. (1999). *Knowledge in a social world*. Oxford University Press.

Gouvea, J., & Passmore, C. (2017). 'Models of' versus 'Models for': Toward an Agent-Based Conception of Modeling in the Science Classroom. *Science and Education*, *26*(1–2), 49–63. https://doi.org/10.1007/S11191-017-9884-4/TABLES/1

Grégis, F. (2015). Can we dispense with the notion of 'true value'in metrology? In *Standardization in Measurement* (pp. 95–108). Routledge.

Guyon, H., Kop, J.-L., Juhel, J., & Falissard, B. (2018). Measurement, ontology, and epistemology: Psychology needs pragmatism-realism. *Theory & Psychology*, *28*(2), 149–171. https://doi.org/10.1177/0959354318761606

Hacking, I. (1999). The social construction of what? . In *The social construction of what?* Harvard University Press.

Hayduk, L. (2014). Seeing perfectly fitting factor models that are causally misspecified: Understanding that close-fitting models can be worse. *Educational and Psychological Measurement*, *74*(6), 905–926.

Heilemann, M. v., Lee, K., & Kury, F. S. (2003). Psychometric Properties of the Spanish Version of the Resilience Scale. *Journal of Nursing Measurement*, *11*(1), 61–72. https://doi.org/10.1891/jnum.11.1.61.52067

Hibberd, F. J. (2019). What is scientific definition? *Journal of Mind and Behavior*, *40*(1). https://psycnet.apa.org/record/2019-31815-002

History and Innovation - What is the Nation's Report Card | NAEP. (n.d.). Retrieved May 15, 2022, from https://nces.ed.gov/nationsreportcard/about/timeline.aspx

Hubley, A. M., & Zumbo, B. D. (2017). *Response Processes in the Context of Validity: Setting the Stage*. 1–12. https://doi.org/10.1007/978-3-319-56129-5_1

Humphry, S. M. (2011). The role of the unit in physics and psychometrics. *Measurement*, *9*(1), 1–24. https://doi.org/10.1080/15366367.2011.558442

Joint Committee for Guides in Metrology (JCGM). (2008). *JCGM 100:2008, Evaluation of measurement data—Guide to the expression of uncertainty in measurement (GUM)*. Sèvres: JCGM.

Jost, J. T., & Gustafson, D. F. (1998). Wittgenstein's Problem and the Methods of Psychology. *Theory & Psychology*, *8*(4), 463–479. https://doi.org/10.1177/0959354398084002

Kane, M. (2008). The Benefits and Limitations of Formality. *Http://Dx.Doi.Org/10.1080/15366360802035562*, *6*(1–2), 101–108. https://doi.org/10.1080/15366360802035562

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological Review*, *95*(2), 163.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (Vol. 111). Chicago University of Chicago Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Lange, M. (2002). Baseball, pessimistic inductions and the turnover fallacy. *Analysis*, *62*(4), 281–285. https://doi.org/10.1093/ANALYS/62.4.281

Lovasz, N., & Slaney, K. L. (2013). What makes a hypothetical construct "hypothetical"? Tracing the origins and uses of the 'hypothetical construct' concept in psychological science. *New Ideas in Psychology*, *31*(1), 22–31. https://doi.org/10.1016/J.NEWIDEAPSYCH.2011.02.005

Lundmann, L., & Villadsen, J. W. (2016). Qualitative variations in personality inventories: subjective understandings of items in a personality inventory. *Qualitative Research in Psychology*, *13*(2), 166–187. https://doi.org/10.1080/14780887.2015.1134737

Luthar, S. S., Cicchetti, D., & Becker, B. (2000). The construct of resilience: a critical evaluation and guidelines for future work. *Child Development*, *71*(3), 543–562. http://www.ncbi.nlm.nih.gov/pubmed/10953923

Maraun, M. D. (2016). Measurement as a Normative Practice: Implications of Wittgenstein's Philosophy for Measurement in Psychology. *Http://Dx.Doi.Org/10.1177/0959354398084001*, *8*(4), 435–461. https://doi.org/10.1177/0959354398084001

Maraun, M. D., & Gabriel, S. M. (2013). Illegitimate concept equating in the partial fusion of construct validation theory and latent variable modeling. *New Ideas in Psychology*, *31*(1), 32–42. https://doi.org/10.1016/j.newideapsych.2011.02.006

Maraun, M. D., & Halpin, P. F. (2008). Manifest and Latent Variates. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1–2), 113–117. https://doi.org/10.1080/15366360802035596

Mari, L. (2013). A quest for the definition of measurement. *Measurement: Journal of the International Measurement Confederation*, *46*(8), 2889–2895. https://doi.org/10.1016/j.measurement.2013.04.039

Mari, L., Carbone, P., & Petri, D. (2012). Measurement fundamentals: A pragmatic view. *IEEE Transactions on Instrumentation and Measurement*, *61*(8), 2107–2115. https://doi.org/10.1109/TIM.2012.2193693

Mari, L., Wilson, M., & Maul, A. (2021). *Measurement across the sciences: developing a shared concept system for measurement*. Springer.

Markus, K. A. (2008). Constructs, Concepts and the Worlds of Possibility: Connecting the Measurement, Manipulation, and Meaning of Variables. *Http://Dx.Doi.Org/10.1080/15366360802035513*, *6*(1–2), 54–77. https://doi.org/10.1080/15366360802035513

Markus, K., & Borsboom, D. (2013). *Frontiers of Test Validity Theory*. Routledge.

Masten, A. S. (2001). Ordinary magic: Resilience processes in development. *American Psychologist*, *56*(3), 227–238. https://doi.org/10.1037/0003-066X.56.3.227

Maul, A. (2013). On the ontology of psychological attributes. *Theory & Psychology*, *23*(6), 752–769. https://doi.org/10.1177/0959354313506273

Maul, A., Mari, L., Torres Irribarra, D., & Wilson, M. (2018a). The quality of measurement results in terms of the structural features of the measurement process. *Measurement*, *116*, 611–620. https://doi.org/10.1016/J.MEASUREMENT.2017.08.046

Maul, A., Mari, L., Torres Irribarra, D., & Wilson, M. (2018b). The quality of measurement results in terms of the structural features of the measurement process. *Measurement*, *116*, 611–620. https://doi.org/10.1016/J.MEASUREMENT.2017.08.046

Maxwell, G. (2009). The ontological status of theoretical entities. *Philosophy of Science: An Historical Anthology*, 451–458.

Mcgrane, J. A. (2015). *Stevens' forgotten crossroads: the divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century*. https://doi.org/10.3389/fpsyg.2015.00431

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383. https://doi.org/10.1111/J.2044-8295.1997.TB02641.X

Michell, J. (1999). Measurement in Psychology: Critical History of a Methodological Concept. *Measurement in Psychology a Critical History of a Methodological Concept*. https://doi.org/10.1017/CBO9780511490040

Michell, J. (2005). The logic of measurement: A realist overview. *Measurement: Journal of the International Measurement Confederation*, *38*(4), 285–294. https://doi.org/10.1016/j.measurement.2005.09.004

Michell, J. (2008). Conjoint Measurement and the Rasch Paradox. *Theory & Psychology*, *18*(1), 119–124. https://doi.org/10.1177/0959354307086926

Monton, B., & Mohler, C. (2021). Constructive Empiricism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University.

National Assessment Governing Board. (2021). Reading Framework for the 2026 National Assessment of Educational Progress. In *NAEP*.

National Commission on Excellence. (1983). *A Nation At Risk: The Imperative For Educational Reform. An Open Letter to the American People. A Report to the Nation and the Secretary of Education.* https://eric.ed.gov/?id=ED226006

Newton-Smith, W. H. (2002). *The rationality of science*. Routledge.

Ogden, C. K., & Richards, I. A. (1925). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism* (Vol. 29). Harcourt, Brace.

Ponnock, A., Muenks, K., Morell, M., Seung Yang, J., Gladstone, J. R., & Wigfield, A. (2020). Grit and conscientiousness: Another jangle fallacy. *Journal of Research in Personality*, *89*, 104021. https://doi.org/10.1016/J.JRP.2020.104021

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Putnam, H. (1987). The many faces of realism . In *The many faces of realism*. Open Court.

Quine, W. v. (1948). On what there is. *The Review of Metaphysics*, *2*(1), 21–38.

Scheffler, I. (1963). The language of education. *Philosophy*, *38*(144).

Schiefer, H. F. (1933). The Compressometer: An Instrument for Evaluating the Thickness, Compressibility and Compressional Resilience of Textiles and Similar Materials. *Textile Research Journal*, *3*(10), 505–513. https://doi.org/10.1177/004051753300301005

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies in History and Philosophy of Science Part A*, *42*(4), 509–524. https://doi.org/10.1016/J.SHPSA.2011.07.001

Slaney, K. L., & Garcia, D. A. (2015). Constructing psychological objects: The rhetoric of constructs. *Journal of Theoretical and Philosophical Psychology*, *35*(4), 244–259. https://doi.org/10.1037/teo0000025

Slaney, K. L., & Racine, T. P. (2011). On the ambiguity of concept use in psychology: Is the concept “Concept” a useful concept? *Journal of Theoretical and Philosophical Psychology*, *31*(2), 73–89. https://doi.org/10.1037/A0022077

Slaney, K. L., & Racine, T. P. (2013). What’s in a name? Psychology’s ever evasive construct. *New Ideas in Psychology*, *31*(1), 4–12. https://doi.org/10.1016/J.NEWIDEAPSYCH.2011.02.003

Smith, B. W., Dalen, J., Wiggins, K., Tooley, E., Christopher, P., & Bernard, J. (2008). The brief resilience scale: Assessing the ability to bounce back. *International Journal of Behavioral Medicine*, *15*(3), 194–200. https://doi.org/10.1080/10705500802222972

Southwick, S. M., Bonanno, G. A., Masten, A. S., Panter-Brick, C., & Yehuda, R. (2014). Resilience definitions, theory, and challenges: interdisciplinary perspectives. *European Journal of Psychotraumatology*, *5*(1). https://doi.org/10.3402/ejpt.v5.25338

Stanford, K. (2021). Underdetermination of Scientific Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University.

Steiner, D., & Bauerlein, M. (2020, October 13). A Feel-Good Report Card Won’t Help Children. *City Journal*. https://www.city-journal.org/naep-proposes-changes-to-reading-tests

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*(2684).

Tal, E. (2013). Old and New Problems in Philosophy of Measurement. *Philosophy Compass*, *8*(12), 1159–1173. https://doi.org/10.1111/PHC3.12089

Tal, E. (2019). Individuating quantities. *Philosophical Studies 2019 176:4*, *176*(4), 853–878. https://doi.org/10.1007/S11098-018-1216-2

Thagard, P. (2007). Coherence, Truth, and the Development of Scientific Knowledge. *Philosophy of Science*, *74*(1), 28–47. https://doi.org/10.1086/520941

van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

van Fraassen, B. C. (2012). Modeling and measurement: The criterion of empirical grounding. *Philosophy of Science*, *79*(5), 773–784.

Werner, E. E. (1995). Resilience in Development. *Current Directions in Psychological Science*, *4*(3), 81–85. http://www.jstor.org/stable/20182335

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Lawrence Erlbaum Associates.

Windle, G. (2011). What is resilience? A review and concept analysis. *Reviews in Clinical Gerontology*, *21*(2), 152–169. https://doi.org/10.1017/S0959259810000420

Windle, G., Bennett, K. M., & Noyes, J. (2011). A methodological review of resilience measurement scales. *Health and Quality of Life Outcomes*, *9*(1), 8. https://doi.org/10.1186/1477-7525-9-8

Wittgenstein, L. (2009). *Major Works: Lugwig Wittgensteing* (First). Harper Collins Publishers.

Wittgenstein, L. (2010). *Philosophical investigations*. John Wiley & Sons.

Xie, Y., Peng, L., Zuo, X., & Li, M. (2016). The Psychometric Evaluation of the Connor-Davidson Resilience Scale Using a Chinese Military Sample. *PloS One*, *11*(2), e0148843. https://doi.org/10.1371/journal.pone.0148843

Zumbo, B. D. (2017). Trending away from routine procedures, toward an Ecologically Informed In Vivo View of Validation Practices. *Measurement: Interdisciplinary Research and Perspectives*, *15*(3–4), 137–139. https://doi.org/10.1080/15366367.2017.1404367

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). *A Methodology for Zumbo's Third Generation DIF Analyses and the Ecology of Item Responding*. https://doi.org/10.1080/15434303.2014.972559

Chapter 4
Empirical methods for investigating the measurand: DIF Modeling

In this chapter I aim to provide a workflow for refining the definition of a measurand via reasoning through statistical and substantive models. The measurand is that which we would like to measure. So a property can become a measurand when it is the target of a particular measurement process. Therefore, the measurand definition may be far more specific than the property definition where the definition of a measurand is specific given capabilities of the instrumentation. When I say reasoning through models, I mean something similar to language dependence in thought. About language-dependent thought, Searle (1995) says, "some thoughts are language-dependent in the sense that an animal could not have that very thought if the animal did not have words or some other linguistic devices for thinking that very thought" (p. 61). This is reminiscent of Fricker's (2007) notion of epistemic injustices (from Chapter 2), namely, hermeneutic injustice when experiences have not been articulated by certain subgroups because those subgroups have not had the opportunity to articulate them (or those very ideas are not considered relevant experiences due to power differences).

Just as we can reason through language, via language, or about language, we can reason through models. Model-dependent reasoning or model-dependence means we have actionable conceptions of what we would like to measure without the model itself. In other words, having a shared model of measurement, to some degree, renders something measurable or hard to interpret. For instance, consider the debate in NAEP reading framework. There lacks a model of the measurand and how it might interact with the particular items. This renders discussions about the measurand difficult because researchers have few referents to discuss where disagreements might be arising or how to coordinate.

Of course, in the previous chapter, I also walked through warnings about thinking (or lack of thinking) becoming too language or model-dependent to the extent the thought or conception cannot be revised (or the model is mistaken as being some sort of exact replica of the phenomena). Wittgenstein called this general conception a case where we might be bewitched by our words. But so too, might we be bewitched by our models. Clearly, it is a delicate balance between avoiding a bewitching and fruitful modeling. My aim here – drawing on a recent study in the measurement of student reading abilities – is to show that terminology from psychometrics and metrology and tools from general(ized) latent variable modeling can help guide conceptual work. While typical latent variable models, in their most charitable interpretation, may be models of item responding (though, this is arguable), they may not be explicitly models of the entity itself – the "factor" in the model.

It is perhaps clear that one challenge with the NAEP reading framework is that the intended use of NAEP is also relatively vague (aside from "monitoring"). This makes it difficult to select among the vast number of potentially causally responsible properties in the act of reading that we *should* measure and the way to define those properties for measurement.[19] In other words, seeking reliability (in the non-psychometric sense and more in the colloquial sense – how much we trust results from the measurement) of measurement results is in part about identifying the primary causal force on the measuring instrument. However, sometimes we need also need to reason given a modeler's intention, what should not be part of the measurand definition or what should be. I aim to use an example throughout this chapter to show how this might look.

4.1 The Empirical Case Study – Measuring Reading

The SUM is a computer-delivered instrument intended to measure four reading strategies which are hypothesized to be attributes of people that vary continuously (see, Arya et al., 2020 for greater detail of the SUM development, piloting, statistical analyses, qualitative analyses, and aims for use) . It is also believed that these strategies can be targeted

[19] Perfetti & Stafura (2014) go so far as to open their paper with the claim "There is no theory of reading, because reading has too many components for a single theory" (p. 22).

for instruction. The four attributes were associated with a particular reading intervention (curriculum) called Collaborative Strategic Reading (CSR; Klingner & Vaughn, 1999). The first attribute of interest is *student abilities or awareness of morphemes in words* named *morphological awareness (*MA*).* This involves looking at words of multiple syllables and breaking down the word into smaller units called morphemes. For instance, the word "illegal" has the prefix *il* which is a morpheme and the single word, *legal*. Recognizing these together that *il* is an affix (prefix) with a particular meaning attached to a word stem (even if a student does not know what a morpheme is) allows a student to understand the word - something is not legal. A second property is what might be termed, *the ability to use context clues to understand a word meaning* (thus termed CC as in context clues). This is a reader's ability to use sentences in which a particular word is embedded to make meaning of that word. The SUM, here, to minimize the effects of other abilities or background knowledge, used made up words as the focal word a student was asked to make meaning of. The third property is perhaps closest to a general reading comprehension property - the *ability to map micro and macro relationships in a text* (MMRT) to key ideas that are prominent or important to the text. In these items, students read sentences or short passages and then were asked to choose among subsequent sentences that were most (or least) related to the focal sentence. Finally, the fourth property was *knowledge of English-Spanish cognates* (COG). The SUM was intended to be used with a multilingual population of students, many of whom spoke Spanish at home. Cognates are words that two languages may share or (of course, there is also the idea of *false* cognates – words that are similar or even the same in two different languages but have different meanings). The motivation for including this property is that students can use cognate knowledge in reading to understand words or even affixes/morphemes.

Hence, the SUM is ultimately and broadly meant to measure *reading strategies* as opposed to comprehension in general. Ideally, each property of students could be used to take relevant action in the classroom. As noted by Afflerbach, Pearson, & Paris (2008), for instance, reading strategies and reading skills are overlapping albeit separate actions or person attributes:

> Reading strategies are deliberate, goal-directed attempts to control and modify the reader's efforts to decode text, understand words, and construct meanings of text. Reading skills are automatic actions that result in decoding and comprehension with speed, efficiency, and fluency and usually occur without awareness of the components or control involved (Afflerbach et al., 2008, p. 368)

In other words, we might see skills as lower level, less intentional forms of knowledge (perhaps we invoke the word *automaticity* when describing a skill), though, one can imagine a strategy that evolves into a skill. Nonetheless, the properties that are the measurands of the SUM are hypothesized to be related but distinct. Drawing on the philosophical discussion of the previous chapter, the aim here is to offer an example about how a local version of epistemic iteration about these properties discussed above and methodological iteration might occur. This is a relatively narrow view of epistemic iteration here in which refining knowledge of properties for the sake of definition and measurement, might be one part of a coordinated and iterative process. From the perspective of ontology, we might admit into our ontology the properties described above and current definitions (or even, descriptions of them) because we believe they can be intervened on and hence affect student everyday experiences of reading (see, for instance, Afflerbach et al., 2008; Arya et al., 2020; Juel & Minden-Cupp, 2000; Klingner & Vaughn, 1999; Perfetti & Stafura, 2014) . These properties shall persist in some form through time and space (a student will be characterized by these properties, in this case, dispositions, even after a test administration or after they are realized). Yet, we can always improve our knowledge, or work to improve our knowledge.

4.1.1 Exploratory Statistics and Methodological Iteration

Elliot (2012) differentiates between epistemic iteration, discussed in the previous chapter – the continuous revision of knowledge claims – and methodological iteration – "a process by which scientists move back and forth between *modes of research*" (Elliot, 2012, p. 378). Elliot defines a mode of research in broad terms, such as moving between constrained hypothesis testing frameworks and exploratory frameworks. Elliot does not exclude using, for instance, different models in service of an exploratory cause or the design of new tools or ways of doing an experiment, for instance as forms of methodological iteration. These two ideas, epistemic and methodological iteration are hypothesized to be connected in the service of different scientific goals. In particular, Elliot posits about methodological iteration:

> First, it can initiate epistemic iteration by helping to provide an initial model, theory, or regularity that can serve as a starting point for subsequent improvement. Second, methodological iteration can equip epistemic iteration by clarifying the nature of scientific problems and suggesting promising ways to revise previous models or theories in response to them. Third, it can stimulate epistemic iteration by helping to identify new problems with existing regularities or models (Elliot, 2012, p. 377; Elliot provides a case study of this process in mRNA research).

In the case of the SUM or instrument construction in general, we might see this in terms of moving between somewhat exploratory work in the form of focus group interviews or even ethnographic work that might involve observing a classroom or just looking at descriptive statistics; more confirmatory work such as using a Rasch model to check to see if items conform to fit analyses; and perhaps returning to some sort of follow-up triangulation approach in which cognitive interviews are performed after the statistical analyses to see if any of the findings from the statistical analyses appear in the cognitive interviews (this was much of the SUM construction process). These moves are instrument-specific. They are, in effect, investigating how people interact with the instrument in different ways. This involves collecting different forms of data (interview data, speech data, item response data) as well as various modes for enquiring about this information. This is perhaps narrower than some intended goals of epistemic iteration, but, nonetheless, seems to follow the same logic. This instrument interaction is the focus of this chapter and the focus is on building a robust plurality of evidence that can support a particular claim about what is measured.

One such set of goals I'm concerned with is the theorizing about, and refining of, definitions of properties of interest for the sake of measurement. How might we learn from data already collected about the properties of interest? Haig (2013, 2005) has proposed that exploratory statistics itself can be a nice starting point for theory generation – either as a "purer" abductive process (hypothesis generation from observing some patterns in data) or as inference to the best explanation (an exploratory process to consider plausible theories of data generation and select among them). However, this may also over admit certain ideas or properties into an ontology without some strong limiting conditions. Here, perhaps, pragmatics can serve as that limiting condition where a property is admitted if it is plausible and reasonably actionable should we have information about it (of course, *plausibility* is an easy term to use but perhaps hard to define). Alternatively, we must still be wary of conflating a model with that which we are modeling (a factor is a term in a factor analysis model and the entity it is supposed to correspond to, if it corresponds to anything, will be different – the factor, at the very least is an idealization) and that even before data is collected or before data is analyzed, there is a model in mind (even if not a stats model).

4.1.2 A Note on Fairness

Some may still wonder about notions of fairness. So often, DIF testing is taken as a matter of checking for fairness of the test among groups where like groups, matched on ability, should have an equal chance of getting an item right (Nisbet, 2019; Nisbet & Shaw, 2019). This is a scoring motivated, rule-based (deontic, almost) approach to fairness in psychometrics. However, in the case of defining the measurand effects what “ability” to match on is. The nature of this focal property and why it is measured will also dictate the argument about test fairness. Some argue that what makes a test fair is if it can be used in service of student learning, which means some tests may direct students on a learning path that differs from other student, violating the typical psychometric approach to treating all “alike” cases as equal (Davies, 2010; Kunnan, 2007; Mislevy et al., 2013; Poehner, 2011). In this case, one can see that how we perform DIF testing would also change. If we include Spanish language knowledge as some sort of an influence quantity or not (and hence part of the measurand definition – such that, e.g. morphological awareness can involve Spanish or English-based made-up words) may depend on how and why students might be reading texts and the decisions someone might make about students.

4.2 Modeling, Measurement, and Measurement Models

To further explain, consider the famed Thurstone's (1940) opening to his paper, "Current Issues in Factor Analysis":

> "Factor analysis is not restricted by assumptions regarding the nature of the factors, whether they be physiological or social, elemental or complex, correlated or uncorrelated. It assumes that a variety of phenomena within the domain are related and that they are determined, at least in part, by a relatively small number of functional unities, or factors…The name for a factor depends on the context, on one's philosophical preferences and manner of speech, and on how much one already knows about the domain to be investigated." (Thurstone, 1940, p. 189)

For the most part, we can see from the perspective of Thurstone, that a generic factor model is not a model of something like human cognition but a tool of, perhaps, discovery. He seems to be warning us at the same time not to read too much into factor model solutions (lamenting that the exploratory nature is not understood). Thurstone seemingly advocates for factor analysis as a first step in a discovery process that is "superseded as quickly as possible by rational formulations in terms of the science involved" (Thurstone, 1940, p. 189). If there is a lesson from Thurstone, it is that any of our latent variable models, whether they be positing continuous "factors" (factor analysis, IRT) or categorical "factors" (latent variable mixture modeling), should be viewed as weak claims about underlying properties and that we should perhaps even avoid naming! That is, these latent variable models, according to Thurstone, are in use due to our lack of knowledge about some set of hypothesized properties, as was stated in Chapter 2 (this position was restated and expanded by Borsboom

(2008)). Said another way, there is a rough assumption that there is some patterning to the mind which is already a model of the mind.

Of course, there are some philosophical difficulties to Thurstone's position, especially given realist accounts of modeling. For instance, the factors are, in effect, only related to *this* (the factor model) model of the mind even, though, clearly, the generic factor model is not meant to be a model of the mind. However, we cannot be completely "theory-less" here, because using the particular observations in estimating a factor model implies that we think that data is relevant for making inferences about the target properties. Thurstone's idea that moving from the model to laboratory experiment as the only direction (as opposed to refining or adding a substantive model) perhaps over-relies on mimicking physics or an idealization of how a trustworthy science advances. For instance, in the earth sciences, it is hard to imagine how moving from a model of plate tectonics (or measuring large features of earth or other planets) to laboratory experiments would be possible. Yet, it would be an odd contention to take that the earth sciences have not been productive in the realm of plate tectonics (of course, it may be *aided* by lab-based experiments about chemistry or physical occurrences).

Alternatively, we might attribute the advance of these sciences to model-based reasoning. Unlike Thurstone, Miyake (2015) argues that advances in "certain propositions are needed in order to extract phenomena from raw data—to "turn data into evidence"" (quotations internal to the quotes Miyake's own; p. 830). In other words, for any generic factor modeler to treat, say, item response data as "evidence" for a measurement claim or claim about discovered phenomena, there must be a set of propositions in which the factor model instantiates. Deviations from the factor model are then more informative. A similar though wider ranging and thorough set of arguments in this line were made by Maraun (1996) and McGrane & Maul (2020). Perhaps we overly rely on off-the-shelf models to direct our thinking, item development, and selection. But perhaps this is less problematic if we heed Thurstone's warnings. These models can be starting points.

Throughout this chapter, I aim to pitch a workflow with the position that the same statistical model can serve different purposes. Off-the-shelf latent variable models cannot do this job alone, though, they are also not worthless. The next stage could simply be a refinement of the model or new way of collecting data (or different data) under different model assumptions (model could be statistical or otherwise – for instance moving from a confirmatory Rasch model to a more exploratory model positing differently structured properties). This has ethical implications in the human sciences – for instance, naming factors or classes imposes the risk of turning them into, in our minds, elements of the world that have a definition. Instead, properties should motivate the models. An interesting exception may be the Rasch model (Rasch, 1960). This is explicitly supposed to be a model of measurement, though, not necessarily a model of cognition (Andrich, 2004; though, it likely should have some correspondence to the structure of cognition to be considered a measurement model).

In the case of something like the SUM, construct maps (see previous chapter) were used as models of cognition, or at least, as models of item response tendencies. In this case, iterating between confirmatory and exploratory models using default, built-for-no-phenomenon latent variable models can lead to a scientific aim of improvement in which a cognitive model keeps the property which we would like to measure as the "truth-maker". A "truth-maker" is that which makes something true. So the state of the world makes a statistical model at least not wrong as opposed to model fit doing the truth-deeming. In that sense, there need be some evaluation of the quality of the interaction of the construct map (Wilson, 2005, 2013)with the generic Rasch model which is meant to be a model of measurement.

In this chapter, I will particularly focus on the repurposing and reconsidering the role of differential item functioning (DIF) modeling[20] and testing in psychometrics. I will then use this to provide an empirical example by working through considering a framework for DIF in terms of connecting what metrologists might call influence quantities and the identification of meaningful subgroups in DIF analyses and how this relates to measurement, models of cognition, and (statistical) modeling practice. Statistical modeling here is aimed at investigating the property and hence the measurand.

Specifically, I aim to repurpose the concepts related to DIF and DIF procedures to use as a model of item responding (Zumbo, 2007; Zumbo et al., 2015). This is perhaps typically seen as only relevant to instrument creation for teaching purposes.[21] However any sort of large-scale, high stakes assessment is not immune to improvement either, and would also benefit from measurand refinement through similar methods. However, the goal will not be

[20] Of course, one could argue that a generic mathematical model (e.g. an equation for DIF testing that applies to all settings) is not really a model at all since there is nothing explicitly being modeled (Giere, 2009; Gouvea & Passmore, 2017).

[21] I'm trying to avoid the terms *formative* and *summative* since it seems that there is nothing in particular that makes an instrument formative or summative – a specific instrument used for formative purposes could be standards aligned in some way and be used for summative purposes. Formative and summative uses then require much work from the teacher outside of the actual test-construction and test administration process (see Nitko, 1995, for an example where an author switches between terms like "summative use" and "formative use" of tests as opposed to, e.g. "summative test")).

4.3 Important Measurement Terms in Statistics, Psychometrics, and Metrology

The aim of this section is really to show that the statistical grounding of psychometrics need not be that different from the grounding of metrology and to then introduce the idea of influence quantities as an umbrella concept linking different forms of measurement (non)invariance in psychometrics, construct irrelevant variance, and perhaps even linking and equating difficulties to a causal notion. Psychometric language may inch forward when concepts from metrology can be incorporated – many psychometric methods using statistical models just have a different semantic structure than that of metrology-recommended uses of statistical tools.[22] The goal is to move from a conception of what we measure as whatever conglomerate of things that cause measurement results to specific reasons.

The language of metrology here is useful because of the specificity in its terminology relative to the terminology in psychometrics – after all, the terminology is older and more developed resulting in the *Vocabulaire international de métrologie* (*International Vocabulary of Metrology;* or the VIM) via The Diplomatic Conference of the Metre in 1875 (discussed in chapter 1 of this dissertation (1.2)). The VIM, in its attempt to define what is measured, can certainly be fallible but it is clear in its attempt to put what is measured and the definition of measurement front and center. Nonetheless, given the specificity of the VIM and the lack of specificity of terms like "latent variables", the language of the VIM, or at

[22] Though, certainly, there is no lack of tooling or thought here in psychometrics such as the tooling in structural equation modeling (SEM) which can incorporate, in principle, many measured properties in causal relation and the very idea of using multi-trait multi-method designs or approaches.

least the intent of the VIM and the field of metrology will be used. In this case, the definition of the measurand and the term is nice for multiple reasons:

1. In the field of statistics, the quantity to be estimated is called the *estimand.* So the *measurand*[23] has an equivalence in language. For instance, if we want to estimate students' morphological awareness, then *a group of students' morphological ability* is the estimand and the measurand in a statistical and/or measurement model. The estimand is what we are interested in when using a statistical model – the quantity to make inferences about. There have been recent calls in psychology and causal inference to define the estimand carefully, for instance (for instance, Lundberg et al., 2021, but also, as central to causal inference in general - e.g. Morgan and Winship, 2015). The measurand is the quantity (or property) one aims to get information about or make inferences about. If one were to use statistical methods to help estimate and correct for what, for now, we'll call errors in a measurement process, the concept of an estimand and a measurand merge. In fact, the history of statistics and measurement are not independent (e.g. chapters 1 and 2 of Stigler, 1986) but are certainly not the same thing, either.
2. Also in the field of statistics, the notion of an *estimate* is the result of constructing (or using) an *estimator*. In this case, one might decide on an estimand and decide how to estimate what the value of that estimand may be. For instance, one may commit to using maximum likelihood for estimating the mean population value of some quantity

[23] The online version of the VIM in its annotation of this ([VIM3] 2.3 Measurand, 2014)([VIM3] 2.3 Measurand, 2014) notes that an older definition used to say, "particular quantity subject to measurement" but that what is subject to measurement may not be what one intends to measure effectively conflating the measurand with the measurement result or the estimand with the estimate.

(student morphological awareness) that one presumes to be normally distributed in the population. The maximum likelihood estimator for parameter μ the normal distribution is:

$$\frac{1}{n}\sum_{i=1}^{n} x_i$$

where n might be the total number of people in a sample, and x_i is the morphological awareness value of student *i*. Passing data through the formula would yield an estimate. The equivalent of an *estimator* in the VIM or metrology may be a *measurement procedure*:

> "Detailed description of a measurement according to one or more measurement principles and to a given measurement method, based on a measurement model and including any calculation to obtain a measurement result" (JCGM, 2012: 2.6).[24]

Note that there may still be an estimator involved but taking measurement as a process to begin well before one has data also means the equivalent of an estimator in measurement has to before one has data.

[24] A convenient replacement of the *Standards's* definition of validity may be also from the VIM in its definition of a *reference measurement procedure* (bolded terms in the document have definitions in the VIM): **"**accepted as providing **measurement results** fit for their intended use in assessing **measurement trueness** of **measured quantity values** obtained from other measurement procedures for **quantities** of the same **kind**, in **calibration**, or in characterizing **reference materials"** (JCGM, 2012; 2.7). This perhaps more eloquently discusses what, I think, the *Standards* are trying to communicate - that specific instruments have specific uses and a measurement process specific to that measurement problem is necessary. Note as opposed to the *Standards* where validity is the hallmark, this is not a hallmark of the VIM, whereas definitions of measurement, measurands, and related are first emphasized given the multiplicity of uses.

3. We can see that an *estimate* from a statistical model maps well to a *measurement result* using a measurement procedure. A measurement result is defined as:

> "set of quantity values being attributed to a measurand together with any other available relevant information" (JCGM, 2012: 2.7)

Where a *quantity value* is defined somewhat circularly by the VIM, but is essentially the result of a measurement - a value (or number) combined with a unit (e.g. 3 lbs or 3 Lexiles). Also, as an estimate attributes a value, the VIM maintains that a measurand has a *true* quantity value. This brings us to the VIM's definition of *measurement*. Again, the terminology from measurement and statistics coincide because these estimates can be aggregation across many occasions (repeatedly measuring something and taking the average; often written in the form $Y = f(x_1,$ $x_2 \dots x_n)$ where Y is the result of multiple measurement outcomes f(x) where f(x) could actually be individual item responses of a student, and y might be the outcome or f(y) could be if some further transformation is performed on Y as well; Here a choice of the function f(.) could be considered the estimator):

> "process of experimentally obtaining one or more **quantity values** that can reasonably be attributed to a **quantity**" (JCGM, 2012; 2.1).

Note how this differs from Steven's definition, in that measurement results are obtained (as opposed to assigned), implying a certain interaction with the real world between the measurement instrument and the *quantity* of interest. While the status of number or quantity values are not necessarily well defined in the VIM it is also not clear how Steven's definition, inducing a certain sort of non-reality-based isomorphism between number and whatever it is that number is assigned to could avoid creating something that’s not a function at all (e.g., measurement results across contexts could map the same value to many different outcomes in different contexts as long as locally held rules are followed – and this would not be due to uncertainty). But as noted in the same section, (JCGM, 2012; 2.1, NOTE 3) – “measurement presupposes a description of the quantity commensurate with the intended use of a measurement result” as well as the measurement procedure. In other words, according to the VIM, one cannot measure without some understanding of what is measured.

4. One concept from the previous chapters is that of an *influence quantity*. The VIM defines the influence quantity as a quantity (or property) “that, in a direct measurement, does not affect the quantity that is actually measured, but affects the relation between the indication and the measurement result”. Here, an influence quantity can be thought of as a confounder in a statistical model, though, in some realms, an influence quantity will only effect the output. The VIM or the GUM, though not specifically statistical documents, are also concerned with uncertainty due to sampling. For instance, in describing why it might use ANOVA, the GUM says it may be concerned in each setting about the influence of an “an “operator effect”, an “instrument effect”, a “laboratory effect”, a “sample effect”, or even a “method effect” in a particular measurement” (Joint Committee for Guides in Metrology (JCGM), 2008). Here, we see a mix of systematic and so-called random sources of uncertainty and each of these effects can factor into a measurement model. In the case of psychometrics, ignoring the semantics and ontological commitments of classical test theory (CTT) for a brief moment, has similar tools such as G-theory (which is not

really a theory), which decomposes variance into different sources of error (Brennan, 2000) and this can be recase in a more general mixed model or IRT framework (Briggs & Wilson, 2007).

Unfortunately, the term *quantity* has been used several times. What the VIM is referring to is some "property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and reference" (JCGM, 2012; 1.1; where a reference can be something like a unit or similar). By way of oversimplification, objects in the world have properties that set them apart from other objects (e.g. Mari et al., 2021) . For instance, a rod has a property of length – which is quantitative in nature in that length has a magnitude and is continuous – but it also has a property of weight. For instance, a rod[25] has a length of 3 inches or a reader has a reading ability of 3 Lexiles.

4.4 Differential Item Functioning (DIF) as a Model of Item Responding

Aekerman (1992), Shealy and Stout (1993), Borsboom, Mellenbergh, & Van Heerden (2002) note that group membership indicators used in differential item functioning (DIF) analyses are often really being used as proxies for the effects we care about. Finding evidence of DIF is a sign that an instrument is sensitive to unintended multidimensionality aside from the property intended to be measured. So, an item response is thought to be a function of the target of measurement and something else that *may* be related to group membership. However, this assumes homogeneity of these causal effects within groups – and this is especially problematic in coarse demographic data often found in large-scale,

[25] One might say that, that what we call a rod is an object that has the property of being a rod, as well (e.g. Orilia & Paolini Paoletti, 2022).

institutional data sets or even just given what sorts of information *can* be meaningfully collected.

Using the language of Gouvea & Passmore (2017) who draw heavily from Giere (2009) and even van Frassen (2012), typical differential item functioning (DIF) testing procedures are *models for* and not necessarily models *of*. Of course, what a model is *of*, or represents, is connected to the purpose of the model, but in the case of DIF testing procedures, the dominant model is a model for a purpose as opposed to of a phenomenon. In psychometric realms, invariance testing procedures in the factor analysis tradition (French & Finch, 2009; Meredith, 1993) or DIF testing procedures in the Item Response Theory tradition (IRT; e.g. Aekerman, 1992; Paek & Wilson, 2011), have mostly been concerned with detection and not necessarily explanation, though, some cursory notions of causes of DIF are often mentioned, such as recasting DIF as (unintended) multidimensionality. This may be in part due to large testing companies like ETS using DIF testing procedures for primarily legalistic reasons, for instance, to flag and remove test items that show evidence of DIF because it may mean that a particular item is easier for one group of students as opposed to another even after matching on some criterion like ability estimates or sum scores. This is explicitly a model for. It is purely descriptive in the sense that causes of DIF are at a different explanatory level than focal causes of item responses. The cognitive causes of DIF are unimportant, to some extent.

In the case of work that has already happened with the SUM, DIF models were used in which each item corresponding to each property was interacted with an indicator of whether a particular student (self-reported) spoke Spanish as a primary language at home. Students are then matched on estimated ability and if the probability of answering a question correctly changes based on whether a student speaks a given language at home, such as Spanish. Said another way, imagine that student 1 is estimated to have the same morphological awareness (MA) as student 2, but student 1 speaks Spanish at home and student 2 does not. If both are administered an MA item and student 1 is estimated to have a higher probability of answering the item correctly *given* their membership in a Spanish speaking group, then the MA item is said to exhibit DIF. It is often customary to remove the item without question (e.g. Borsboom, 2002). However, this need not be the case. In the SUM workflow, an item was removed or flagged only *if* a potential reason for DIF could be identified. In a sense, another property was detected with DIF but this property has to be named.

To formalize, using notation close to Borsboom et. al (2002), measurement invariance can be described by the scenario, that for person *i*, item *j*, and a value on latent trait (property) *T*, with a selection/group membership variable *v*, where P(.) is the probability of a response to item *X*,

$$P\left(X_{ij} = x_{ij} \middle| T = t_i, V_i = v_i\right) = P\left(X_{ij} = x_{ij} \middle| T = t_i\right)$$

EQUATION 1

Non-invariance occurs when:

$$P\left(X_{ij} = x_{ij} \middle| T = t_i, V_i = v_i\right) \neq \ P\left(X_{ij} = x_{ij} \middle| T = t_i\right)$$

EQUATION 2

That is, in equation 1, the probability of a response is the same independent of a group variable *V* (invariant) and in equation 2, the probabilities are not the same. However, there is a pernicious problem that may occur, especially in the realm of self-report instruments, where items may invoke a survey respondent to use their own frame of reference (if asked if I work hard, I might consider what working hard means relative to others in my profession as opposed to other professions). In that case, we may find evidence of differential item functioning across different schools or classrooms (or similar). The *model for* approach adopts a model that, for lack of a better term, “gets the job done” in ensuring the equality in Eq.1 holds given the item set administered to students. In this sense, models are thought only as instruments to fulfill an end goal (Keller, 2000). If traditional DIF testing is a model of anything, it is a model of a particular testing situation (Zumbo, 2007) or a model of data, but is not necessarily a model of cognition when demographic grouping variables are used and items are simply removed when DIF is detected.

Zumbo (2007) acknowledges that even in testing situations at large testing organizations, some may worry about reasons for differential item functioning. Zumbo (2007) quotes Angoff (1993):

> The “why” concerns can be clearly seen in Angoff (1993) when he wrote about long-standing Educational Testing Service DIF work: “It has been reported by test developers that they are often confronted by DIF results that they cannot understand; and no amount of deliberation seems to help explain why some perfectly reasonable items have large DIF values” ”(p. 19).

It will be my argument here, by shifting to a *models of* framework given a modeler’s specific purpose *for* modeling, the realm of measurement invariance research in psychometrics can be used to consider and iterate on definitions of the measurand.

4.4.1 The ordinariness of DIF and measurement invariance research

A typical workflow in educational or psychological measurement might involve fitting some sort of latent variable model to data. If adequate fit is found (adequate by various criteria) of items and/or the initial model, then a researcher might go on to do some sort of

DIF or measurement invariance testing. The goals seem diverse in this second phase of testing, but they might be said to use to avoid bias in a statistical sense or an ethical sense against certain groups of people. Mellenbergh (1994) subsumes the DIF model in an IRT context under the umbrella of what he calls, generalized linear item response theory. Of course, these are not models of phenomena on their own. A similar framework was presented thoroughly in de Boeck and Wilson (2004) in what they call, Explanatory IRT (EIRT). Given a latent property (see chapter 3 for discussion of the coherence of latent variables), Mellengbergh expresses that the generalized item response theory model can be expressed (with a slight change in notation) as, in non-matrix form:

$$\boldsymbol{g(\tau_{ij}) = b_j + a_j t_i + c_{1j} z_{1i} + \cdots + c_{pj} Z_{pi}}$$

EQUATION 3

In the model above $b_j, a_j,$ and c_{pj} are parameters related to (or of) item j, where b_j is effectively the item difficult in IRT terms, or the intercept(s), and *a* could be a discrimination parameter as introduced in chapter 1 and t_i then is treated or modeled as a latent variable and refers to the trait or property of interest for subject *i.* Finally, *z* are subject *i's* scored item response (here, consider items dichotomously scored) and *c* are any number of item features such as item word count, type of item (for instance, a dummy indicator for whether the item might be of a "fill-in-the-blank" or "cloze" format), or the relation between an item and the human property it is associated with. The function g(.) is a link function of some sort – most commonly the logit link. If we only have item intercepts and person abilities, then the model reduces to, effectively a 1PL or 2PL model (depending on the value of *a* – if *a* is not subscripted or is fixed to a value of 1, then we get something like the Rasch or 1PL model*)*:

$$\boldsymbol{g(\tau_{ij}) = b_j + a_j t_i}$$

EQUATION 4

We might instead express this as an IRT model in traditional notation:

$$g(\eta_{ij}) = \theta_i - \delta_j$$

EQUATION 5

Where θ_i is a random latent variable that is an estimate of person ability and δ_j is the item difficulty. We can switch signs and δ becomes item easiness to match Mellenbergh. This is no different than regression or linear model notation, only, variable values for independent variables may be estimated from the presently collected data. However, as Bollen (2002) notes, even in a standard regression we might consider residuals in a standard linear regression random latent variables.

Finally, we might express a differential item functioning model, again, using the notation of Mellenbergh.

$$\boldsymbol{g(\tau_{ij}) = b_j + a_j t_i + c_{1j} z_{1i} + c_{2j} t_i z_{1i}}$$

EQUATION 6

Here, *z* can be an observed or latent variable in the model. Usually, this will be some indicator of group membership. For instance, if we have a group of students that identifies as Spanish speaking, we might code Spanish speakers as *z =1*, and 0 otherwise. In this way, we can see that the parameter, c_2 could also be of interest and be a measurand itself. We can conceptualize DIF, syntactically, as an interaction term between a person's group membership or some other variable and an item, which is expressed in typical IRT notation (following the model of Paek & Wilson, 2011) as:

$$g(\eta_{ij}) = \theta_i - \delta_j + \gamma_j Z_i$$

EQUATION 7

Where γ_j is the difference in item difficulty given group membership Z of person *i*. Below, I will primarily refer to the variable, *Z*, for DIF testing as a “grouping variable” and the quantity of primary interest, θ_i as the measurand. In equation 6, this could be more easily seen as detecting DIF when the item characteristic functions for a single item but for each of the two groups are not parallel. DIF is one special case of including covariates, meaning there is no reason DIF testing should have a particular name except due to its special role it serves. We do not need to stay connected to that purpose, as important as it may be.

Admittedly, in the case of the SUM, a somewhat arbitrary significance value was used (adjusted for multiple testing) for flagging DIF, however, items were further analyzed and investigated via concept and item analysis in terms of words in items that may have had Latinate (or non-Latinate) morphemes. The logic of DIF detection and what to do with the results inherits meaning from a tacit model of item responding – the process of reading involves making use of micro-level information about text, in the words of Kintsch and van Dijk (1978).

This leads to an interesting conundrum in terms of defining properties such as morphological awareness and cognate knowledge that the SUM measures. For a student to have morphemic knowledge, this may be the same knowledge as cognate knowledge, though, at times, they may be knowledge used separately. In this way, one must ask about the definition of the measurand or how to know when cognate knowledge and morphemic knowledge is separate. More specifically, how would we refine the definition of the measurand(s)? One way is to also account for "influence quantities" in a measurement system.

4.4.2 Causal Accounts of DIF

Sometimes it is said, though a bit inelegantly, that we use item responses to measure the property of interest. The only way for the coherent view of this is that there is something that effects the instruments we use for measurement (Borsboom et al., 2004; Giordani & Mari, 2012; Mari, 2013; Markus & Borsboom, 2013) and I would like to attribute the primary cause to the property of interest. We see this notion of causality as well in the very naming of the multiple-indicator multiple-causes model (MIMIC) for DIF testing in factor analytic contexts. This would seem to be a basic tenet of the realism posited in chapter 3. This also means that, while person abilities or *t* in equation 4 are usually of interest in item response theory contexts, we may be interested in the relation between *t* and other focal properties. While this is common in structural equation modeling (SEM), it is unclear if the tenets of thinking about focal and non-focal causes, or treatment effects, are embedded in SEM usually. Regardless, the aim here is to repurpose DIF for inductive inference about the limits, boundaries, and, ultimately, definition of a primary property as well as other properties that are necessary for use in correction of ability estimates in an IRT model.

Given this discussion, we can see also that there is causal implication for DIF. That is, variables used in a DIF detection model should also have causal interpretations and need not only be observed variables, in the psychometric or statistical sense. In fact, it is unlikely that they can be observed variables in the statistical sense because of problems mentioned in chapter 3 – even rudimentary measurements of observed variables can be recast as latent variables in a different statistical model involving repeated measurement (or sampling).

4.4.3 Semantics of Causal Account of DIF

Given two different interpretations of latent variables, though, the causal account may be very different. Holland (1990) introduced the semantic problem in IRT modeling, namely whether an IRT model should be interpreted with the *stochastic subject's interpretation* or the *repeated (or random) sampling interpretation.* If we accept, as is common in the IRT literature, the stochastic subject's interpretation, this means that the probability of a response to an item in a given category (for instance, either correct or incorrect), can be interpreted at the within person level. In the frequentist view, if I am estimated to have a 60% chance of answering an item correctly, I should answer the item correctly 60% of the time in a fictitious scenario of repeated test taking and brainwashing. Alternatively, the repeated sampling or random subjects view, the 60% instead indicates how many people with the same value of a given property will get the item correct. If 100 people with the same level of reading comprehension answer the same question that I do, 60 of them will get the item correct. This means that the variable is not random at the person level in this case, it is only random at the group level. This seems to have interesting implications for differential item functioning. If I adhere to the random sampling perspective along with a between person perspective of the latent variable, this implies that the property under measurement is also at the group level and has a black box interpretation –the between person account does not necessarily hold at the individual level and the structure of the attribute does not hold at the individual level. Thus, there is not necessarily a cognitive explanation for my having the same value of some property value as someone else (or more precisely, I could have the value for very different reasons than somebody else) and the quantitative structure of the attribute at the individual level is not necessarily justified (by quantitative, I mean that it occurs on a continuum as in chapter 3's construct map).

In the between person account, explaining differential item functioning may not be cognitively interesting, since there could be many ways a given non-focal property affects the item response process since, at the individual level, the between person model does not posit a measurand at the person level. Instead, it yields differential item response probabilities within the grouping property such as "Spanish speakers at home" which is also a group level variable and has no cognitive content. It is not clear in this interpretation how one *explains* DIF unless there is a strong manipulable variable at the group level as well (e.g., took Spanish language classes or not). If it is coherent to explain how a grouping demographic of interest causes a change in item responses matching on ability level, the explanation would also have to be at the group level – as opposed to the individual student cognition level. Then, the number of cognitive explanations and, hence, the number of causal forces becomes numerous. For instance, if it is another person property such as a cognitive ability (e.g. Spanish speaking ability which is measured as a continuous attribute), then this may be interpreted as the average causal effect of Spanish speaking ability on the item response for each group of the measurand values (θ in equation 6, above).

The explanation, though, better termed a description, is something simple – "had the values of students with cognitive ability (θ_i) had a higher cognitive ability and that same group with cognitive ability Spanish speaking ability been higher, they would have had a higher chance of answering correctly an item with a DIF effect favoring those with lower Spanish speaking abilities" (I'll call this **DIF Statement 1**). Since, as noted by Borsboom et. al (2003), the between person version of the model only attributes between person claims, there are many possible explanations among individuals for why DIF Statement 1 holds but this is not relevant to the IRT model. This expresses conditional probability differences, where probabilities are expressed as the left side of the equality in Equations 1 and 2, above. Hence, these are population differences, where populations might be defined as given levels of the measurand and grouping variable (e.g., all people with reading comprehension ability .3 logits and Spanish speaking ability .5 logits). When the grouping variable is a something more like a demographic (e.g., whether you are from the cities of Los Angeles or Denver or neither), the cognitive content becomes even less clear. In this case, it is a causal model, but the efficacy of the interpretation of a descriptive variable that does not hold much cognitive value, becomes even harder to explain – it must become something like, "had the students with cognitive ability (θ) had a higher cognitive ability and been from Los Angeles instead of Denver, they would have had a higher chance of answering an item correctly with a DIF effect favoring those from Denver" (DIF Statement 2). This also suggests that there is no reason to believe the DIF effect to be constant within populations, and thus, two people with different cognitive abilities cannot be considered to have within person similarities in the structure of the attribute of interest. This has interesting implications for growth or learning progressions where the metaphor "growth" is likely doing more work than it should without justifying why an increase in scores or estimates warrants discussing quantitative change as opposed to qualitative changes.[26] The above (and below) concerns are really of no matter if one is merely simply intending to tag items to remove which show evidence of DIF so they can be removed.

Alternatively, if we accept a stochastic subject's interpretation along with a within person interpretation of the item response model, then the causes of certain item responses are the same given the same ability estimate. In this case, the between-person model and within person model may match. According to Borsboom et al. (2003) , the stochastic person's assumption invokes a local homogeneity assumption – that is, the between-person and within person structure of the attributes are the same, however, this also is a stronger assumption than the between person model view. In this way, when we use a focal grouping variable for differential item functioning analysis (DIF), then the plausibility of a cognitive explanation for differential item functioning becomes more interesting as the cognitive process involved for each person within the stochastic subject's interpretation is the same. Ideally, then, explaining DIF (e.g. Zumbo, 2007, 2017), becomes a matter of explaining individual cognitive processes. According to Borsboom et. al (2003), the within person, stochastic persons view, while tempting, requires even more thought experimentation. One could use the time dimension or repeatedly sampling within persons to justify the within person account of probability. In this case, the measurement model may become a cognitive

[26] Perhaps this is a result of lingering remnants from operationalist Classical Test Theory (CTT) paradigm. If a true score is not a construct score (as noted by Lord and Novick, 1968)Lord and Novick, 1968), but is simply the expected score ($E[X]$) on a test over successive brain washings, then any change in the positive or negative direction for a true score is, by definition, growth, or increase in the true score. There is no cognitive content posited.

account at the individual level where between person claims invoking the counterfactual listed in the preceding paragraphs for the repeated sampling perspective leads to claims where quantitative differences are indeed, quantitative differences of some sort in the property of interest at the individual level (e.g. Maraun & Gabriel, 2013). In this case, there may be options to treat the grouping variable with a between or within person view. However, the actual effect may have become most meaningful for understanding the DIF effect at a within person level. That is, the DIF effect is most coherent as an effect on individual cognitive processes. Ideally, the grouping variable is also treated or assumed to have a within person effect, though, the concept of non-uniform differential item functioning would seem to imply that there are differing cognitive effects given ability level of a student – though, one wonders is this justifies a within person conception of the measurand, or, if non-uniform differential item functioning, is justified as a differential effect of the "grouping variable."

The point of the section above is to note that much of the work below will assume a stochastic subject's view. However, this is an assumption of sorts, part of the modeling process. Nonetheless, this influences the nature of the explanation provided.

4.4.4 DIF as the Result of Influence Quantities (or Properties)

While a within person conception of IRT is often invoked, it is not an innocent interpretation, and this has implications for DIF interpretation and implementation. But it also has implications for the notion of uncertainty. In previous chapters, the *GUM* was introduced as a guide to categorizing types of and accounting for uncertainty in measurement. However, this has been mostly used in the context of the physical sciences, and uncertainty in social sciences is often from the perspective of statistical considerations, namely, sampling variability (e.g. Rigdon et al., 2019) [27] The *GUM* terminology and VIM terminology around, for instance, instrumental uncertainty vs uncertainty about the measurand is important and can serve as guidance for deciding what sources of uncertainty one is trying to account for. Instead of grouping all uncertainty into the error term, identifying sources of uncertainty in a measurement situation can allow us to refine the model.

One way to decrease uncertainty is accounting for external properties that influence measurement results. The GUM states, "Incomplete knowledge of influence quantities and their effects can often contribute significantly to the uncertainty of the result of a measurement" (JCGM 100:2008, 2008, D.5.3, p. 51). Tal (2019), Sherry (2011), van Fraassen (2012), and Maul et al. (2018) present versions of this sentiment in which measurement is a model-based enterprise. Confidence in measurement results is bolstered by modeling of the measurement process or background knowledge in general. In other words, having a model (or multiple models) of the measurand (could be a definition of the measurand, a picture, metaphor, or even story that serve as models), the way the measurand interacts with the measuring instrument based on this model bolsters confidence in the measurement result and reduces uncertainty. Additionally, as we learn more about the behavior of a measurand, we may know about or invoke more quantities (or properties) in our model of the measurement process. Here, we see where representational models are necessary in addition to the purpose.

[27] Admittedly, I am ignoring discussions of how a measurement result can come to be and how it is evaluated.

To merge with metrology, we can see the properties that cause DIF are influence quantities (or properties; Mari et al., 2021). In metrology, influence quantities not accounted for can increase uncertainty. Additionally, the lack of knowledge about influence properties or not accounting for influence properties can increase uncertainty in a measurement result.

However, this model-based account of measurement should begin to sound like a nomological network from Cronbach and Meehl (1955). A problem with the nomological network perspective in education and psychology is that it focuses on meaning instead of reference of a term such as "*morphological awareness*" and does not requires a definition of the term itself necessarily. As mentioned previously in this dissertation and other papers (Borsboom, 2005; Lovasz & Slaney, 2013), the nomological network of the Cronbach and Meehl variety requires that defining terms "means to set forth the laws in which it occurs" and requires that relations involve at least "some observables" (Cronbach & Meehl, 1955, p. 290). This distinction between observables and unobservable was described as something of a false dichotomy in chapters 2 and 3, however.

In the following, I aim to separate out a nomological network conception with a definition of the measurand *and* the concept of influence quantities which need not be entered into a law like relation for definition. Finally, then, influence quantities or properties can be seen as no different from primaries of focal interest, just the influence quantity is not the primary of focal interest.

4.4.5 (Optional) Connecting DIF testing to item writing

When DIF testing is used as simply to acquiesce a cultural norm or get over a hurdle, then some agency of the modeler is lost (perhaps intentionally) in the modeling process to legalistic or rule-following. However, focusing on individual item by subgroup interactions, as in DIF testing, provides a powerful tool to refine definitions and our models. They allows us to reason from item design to cognitive processes in cases where items are designed with very specific cognitive processes in mind. In other words, reasoning from the way an item is designed to elicit the measurand allows us to think about ways in may cause a person to use other knowledge, skills, or abilities that are not intended to be measured. For instance, consider an item meant to elicit evidence of reading comprehension ability. Reading comprehension itself is an ill-defined property as discussed in chapter 3. However, what it might involve is a reading passage and a set of test questions about those reading passages. In NAEP reading, there are examples in which students are provided instructions about what will be expected of them, and then provided a reading passage of some sort, and questions about the passage – NAEP calls these items "comprehension items" (e.g. National Assessment Governing Board, 2021, p. 99). To consider these items as *indicators* of reading comprehension ability, there must be some reason that a subset of items (or the universe of items) are the ones to use. Not all perspectives, of course, see a causal perspective quite like this. One perspective is behavioral domain theory (BDT). A thorough commentary on BDT is provided in Markus & Borsboom (Markus & Borsboom, 2013). However, considering BDT allows for realizing the central role of causality in item writing.

BDT effectively states that student domain scores are estimated from test scores or response patterns, where a behavior domain is a common set of behaviors related to one attribute. In BDT, though, latent variables or factors are interpreted as "inductive summaries" and "suggest an opening to interpret such a latent variable as an abstraction created from the items in a domain, rather than as a common cause of the item responses" (Markus & Borsboom, 2013, p. 57). This is not exactly formative measurement model because BDT is applicable to a universe of items (though, I suppose, one could say the infinite set of items defining the behavior domain and hence the attribute is some sort of operationalism or formative perspective) – a nonfinite set of items that makeup the behavior domain. In this case, items are selected based on shared characteristics as opposed to causal conceptions connecting the item response to the within person property of interest and the common characteristics justify their selection from or inclusion in the behavior domain (Markus & Borsboom, 2013; McDonald, 2003). McDonald (2003) argues that items are selected or constructed not on the basis of a causal theory but of a semantic-psychological theory:

> "Investigators do not operate a common-cause notion in applications of common factor/item response models. Rather, they write or add them to a given [item] set, to be "of the same kind, " in the sense that the items share a common property with each other" (McDonald, 2003, p. 222).

This has the interesting consequence, according to McDonald (2003, p. 222), that "alternative common properties of alternative item domains" meaning that something like testing for DIF would seem to be testing for something akin to a behavior domain that differs between subpopulations and this begins to feel a bit like operationalism. Additionally, this leads to a somewhat different conception of differential item functioning. I agree with Markus and Borsboom (2013) that it becomes hard to justify the non-causal account when considering the construction of items. Items share many characteristics that are not intended to be part of the behavior domain and are not the primary elicitor of the intended response process. Instead, there are core features of items that are intended to elicit a particular response process. This allows us to construct many different types of items that may be used to measure the same attribute and implies that DIF is caused by item characteristics evoking or causing unintended response processes conditional on something to do with student group membership. Hence, *varying* features of items causes, potentially, different response processes. For instance, Roussos and Stout (1996) state:

> "“Model for DIF”; means some way of linking substantive item characteristics (such as descriptive content, solution strategies required, superficial features, cognitive processes used, and so forth) to whether or to how much DIF will be displayed by an item for particular subgroups of the test-taking population" (p. 356).

Note the language of "model for" as above – this model has a particular purpose. The authors argue that some of this *model for* DIF enters into the process of item writing itself. However, there's a second stage in the Roussos and Stout methodology that brings in a purpose built, more representational model in which they model items hypothesized to exhibit DIF as a multidimensional model in which causes of DIF are what they call, non-focal dimensions.

4.5 Using data modeling to refine definitions of the measurand – a case study

In his brief paper, Mari (2006) provides a cogent model for considering the definition of a measurand. As mentioned in the introduction to this chapter, the definition of the measurand is not quite the same as the definition of a property. Models are abstractions that help us reason via simplification. Mari provides an example of measuring the length of a rod. We might ignore the effects of gravity on the rod but we may also ignore the effects of temperature since the rod may lengthen at higher temperatures (or even change its shape).

Recall from chapter 2 of this dissertation, that Kolen and Brennan stated that in social science we "almost always measure at least some different constructs even if they have similar names" (p. 488). This misconstrues the process of learning about properties through measurement and, at the same time, refining our instruments and definitions. Similarly, even in something as simple as measuring the length of a rod, where the definition of the measurand is something like, "length of rod x", this would not take into account temperature. In this scenario, temperature is maintained as a "hidden variable" in the definition of the measurand" (Mari, 2006, p.2) and this increases uncertainty overall.

Alternatively, we could decrease uncertainty by including the definition of the measurand explicitly by defining the measurand as the length of a rod at a given temperature. Of course, now we have to account for uncertainty related to measurement of temperature but we might be more confident in the causal effect of length on the instrument reading. In this case, temperature is an influence quantity. Additionally, the intentionality of specifying a model of the length of the rod and temperature and the influence on the measurement process helps demarcate temperature from the measurand.

Now consider the SUM discussed above. In the original paper, DIF analyses were used to flag items. Students reported whether they spoke Spanish at home (Arya et al., 2020). We might call this an observed subgroup. In typical DIF testing procedures as discussed above, items are removed to account for *whatever* it may be that causes DIF. In the SUM, items were removed or altered if DIF results could be explained. However, removing items can be costly. Additionally, in the case of the SUM, whether students speak Spanish at home or not, is certainly not at the same cognitive level as student morphological awareness abilities which describe micro level cognitive processes. Alternatively, whether a student speaks Spanish at home is not *necessarily* a measure or indication of cognitive processing. Alternatively, that seems to be the implication of the model of morphological awareness –

students filter word meanings through a sieve of morphemic knowledge. Additionally, as discussed in Arya et. al (2020), some items involved Latinate morphemes that could be aided by Spanish language knowledge. In other words, there is a similar scenario to the rod length scenario above in which we can choose to leave temperature as a "latent" source of uncertainty unaccounted for or include it. In this case, something like, "Spanish language vocabulary knowledge" might be relevant for making instructional decision. Spanish language vocabulary knowledge could be seen as an influence quantity in the other dimensions, or at least, always present. However, how we might even begin accounting for it is tough. Perhaps a first step would be understanding it from the perspective of quantity individuation and exploratory data analysis.

In the initial statistical work on the SUM to understand whether items generated student response processes that could yield data that conformed to measurement, the data was compared to the Rasch model. The model of responding was articulated via construct maps. The extent to which data deviated or conformed to the Rasch model given best estimates of the difficulty of items, was the extent to which we could say items were worthy for use in measurement. Invariant comparison was the primary motivation of this aspect of analysis. This is a confirmatory stage, or at least, relatively confirmatory (for instance, item difficulties were not tested via ordinal constraints, about hypothesized difficulties before data was collected and hence could take any value). Following, this analysis was an exploratory phase using DIF analysis as discussed above. This was more exploratory than the previous phase. Using equation 7 above, items were interacted with *group* membership and flagged for DIF if the item was above a certain value (e.g. following ETS-based DIF values for flagging if an item changes a probability of response by a given amount). Next, items were not removed unless a reason for the DIF could be identified.

The point of the next phase is not to critique but to refine. To some extent, though perhaps tacitly, this was acknowledged by Borsboom et al. (2002) when they described different kinds of DIF. Borsboom et. al (2002) propose that one might be able to find a scaling method such that the distribution of one group may be shifted over and without removing items even when DIF is found due to what they call, relative invariance. This may seem like a strange idea until one considers that we deal with this sort of issue frequently. For instance, we are able to think about the relative value of $1 U.S. based on what it can purchase in different places. For instance, $1 may purchase more in one geographic location than another and it is not uncommon to find comparisons of purchasing power.

This transformation, if it can be achieved, would be powerful. For one, it may save some items from being removed or changed. Perhaps more interestingly, in trying to understand the transformation to be achieved to bring the item response probabilities to equality, a deeper understanding of the response process involved may result. In particular, what is the influence quantity that is shifting the distributions? However, yet another problem remains: How might we decide if a group is homogenous to   some sort of transformation?

Borsboom and colleagues provide an example of height measurements via self-report by asking respondents questions like, “can you reach the top of the bookshelf?” or similar. In this scenario if there are two groups of people, and one group is, on average, shorter than another, this item might show evidence of DIF. For example, imagine students in grade 5 and grade 12 answering in a binary (yes, no) way to the statement proposed by Borsboom et. al. – “I would do well on a basketball team” (Borsboom et al. 2002, p. 435). In this scenario, if we were to match on height (either as known or as estimated from other questions), students in 5th grade have a higher probability affirmatively answering the question given same heights, say, of 5’10” (approximately, 178 cm). If DIF is detected, the item might be removed. However, Borsboom argues, some transformation might be acceptable (such as matching distributions via scaling means within group) where heights are now compared within group, specifically, for that item. If a transformation can be found (for instance, by mean centering height distributions in each group), then the item need not be removed, though, the scale of the item response is shifted such that comparisons are to be made within group.

The caveat is that this transformation, as noted by Borsboom et al. (2002) is contingent primarily on the response process. The height example is quite contrived, but provides an example where the response process is known – the basketball item is known to be non-invariant absent the transformation because students in 5th grade are shorter and compare themselves, potentially, to other 5th graders. The frame of reference of the respondent is known. Of, course, this raises the question about whether this counts as *measurement* at all. In some sense, I would argue it is questionable since, intersubjectively, the results depend on who answers the question. All the same, it provides an interesting discussion into understanding the property of interest via empirical means – and hence, trying to better understand the measurand (and model the response process) when applied across groups, unless, of course, there’s a transformation – though, what is communicated via the measurement result is not a height in absolute terms.

Though, Borsboom et al. (2002) conclude that this transformation methodology does not provide guidance for academic tests (reading, math, etc), it does seem to provide the base for questions about these academic test response processes, using DIF as modeling a tool for investigating the response process (and hence, the measurand). Additionally, though Borsboom et al. also argue that absolute and relative height do not imply multidimensionality (instead, it can be considered a frame of reference problem), I would argue that there are certainly other quantities that influence the response differences that are not the intended to target of the instrument – the relative vs absolute measurement invariance problem can then be seen as a definition of the measurand problem (should height be measured among certain age groups?). Here, we have some initial tooling or thinking for not leaving the unmeasured influence quantity unarticulated (the unspoken of temperature in the rod measuring case). Yet, this analysis of requires knowing that the groups used for DIF analysis are homogenous enough to scale their distributions of values on some property. The question of matching variable (typically a group membership indicator) homogeneity is thus in question. That is, in the case with height above, there is some tacit assumption, that for the case of 5th graders, that the response process is the same for all 5th graders. However, this is an assumption and one that can be tested.

This might have important impacts in the classroom, as well, for academic tests. Assuming a test or instrument will be used to make teaching decisions, effectively contextualizing these teaching decisions via local (or conditional) knowledge of the measurand and the way the property interacts with the test or instrument in general might be of important use. From the perspective of test writing, this could dictate content as well.

Typically, DIF testing is used to remove items to treat all students as equal. However, in the case of the SUM there may be instances where DIF testing, may be useful for identifying student strengths though they are not part of the property of interest. For instance, bilingualism or background knowledge could be seen as an influence quantity in the SUM but also something that might be useful to know about a student for instruction (e.g. Hopewell, 2011; Ramirez, 2000; van Assche et al., 2009). Hence, this implies a theory of the measurand.

4.5.1 Spanish language in the SUM

In the SUM analysis, information about test-takers' language(s) spoken at home were collected. Here, for the case study, we will focus on speakers of primarily Spanish at home. Consider the item stem from the SUM (item – QMA13):

> "The book covers were heterocoloreous. What does heterocoloreous mean?" (the correct answer is many colors)

This item was used as part of the instrument for measuring a student's morphological awareness. We can define the measurand using the definition of morphological awareness above but attribute it to the student with a minimal (quantified?) set of distractions and little background knowledge (this is, of course, relatively inadequate as a definition of a measurand). In this case study, I will focus on exploring the relationship between Spanish language knowledge, the particular cognition involved in answering items and what sorts of item feature invoke particular forms of Spanish language knowledge. This is akin to investigating how the instrument may interact with different properties in different environments.

If we were to take a psychometric lens, this would be akin to explaining sources of differential item functioning. This is part of quantity or property indistinguishability. For instance, do we expect Spanish language knowledge to be distinct from the properties posited by the SUM? What, exactly, is the cognition involved? What elements of items are causally responsible (see, section 4.6.4) for eliciting the responses? I hope the methodology developed below may be of some use.

4.6 Methods: Using Mixture Modelling to Make Meaning of DIF

The proposed workflow in this project starts with a scenario where one has found evidence of DIF (say, via an IRT model). This can be considered step 0 – with differential item functioning, we know the two groups have different distribution. The goal of this process is to try to understand the appropriateness of the "observed group" indicator for DIF analyses or use with relative DIF, definition of the measurand such that one can consider "how much" of another property might need be modeled or is present in items, or even just to do work explaining DIF to articulate the phenomena involved in the item response process above and beyond the measurand (or to combine with the measurand). In other words, the process below is meant to be a hypothesis generating step about phenomena.

The goals of the steps below, are not meant to be set in stone, nor is it original (Zumbo et al., 2015). The aim is to promote a non-theatrical[28] use of psychometric models such that they become a model for an epistemic aim, such as inferential and inductive catalysts. However, the power of the inference is only as good as the model or definition of the measurand and their connection to the property observations. In other words, part of the aim is to reveal that measurement and DIF analyses need not be a black box. This is not meant to be a prescriptive statistical methodology. In fact, further work is needed about investigating statistical properties of, for instance, using the same data multiple times, the ability to use mixture modeling when splitting data into subgroups to make inferences, and

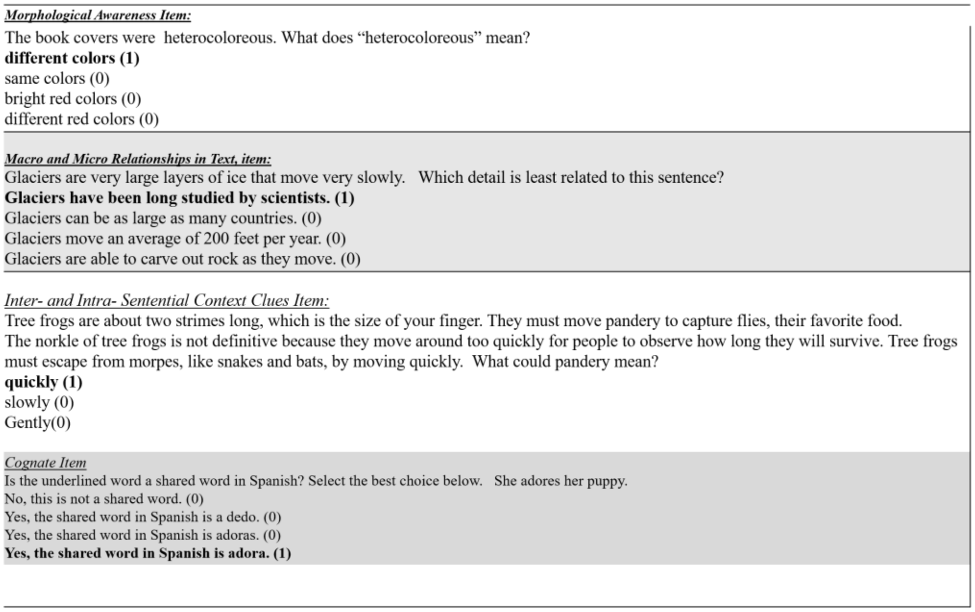
[28] I say theatrical uses to contrast with modeling with purpose. I see a theatrical model as a way to get past the hurdles of acceptance set out by particular audiences.

differences in using, say, mixture modeling vs modeling distributions' shape and scale parameters.

4.6.1 The SUM Case Study – Methods

Step 0:

For the case of the SUM a selection of example items appears in Figure 4.1 below.



Morphological Awareness Item:

The book covers were heterocoloreous. What does "heterocoloreous" mean?
different colors (1)
same colors (0)
bright red colors (0)
different red colors (0)

Macro and Micro Relationships in Text, item:

Glaciers are very large layers of ice that move very slowly. Which detail is least related to this sentence?
Glaciers have been long studied by scientists. (1)
Glaciers can be as large as many countries. (0)
Glaciers move an average of 200 feet per year. (0)
Glaciers are able to carve out rock as they move. (0)

Inter- and Intra- Sentential Context Clues Item:

Tree frogs are about two strimes long, which is the size of your finger. They must move pandery to capture flies, their favorite food. The norkle of tree frogs is not definitive because they move around too quickly for people to observe how long they will survive. Tree frogs must escape from morpes, like snakes and bats, by moving quickly. What could pandery mean?
quickly (1)
slowly (0)
Gently(0)

Cognate Item

Is the underlined word a shared word in Spanish? Select the best choice below. She adores her puppy.
No, this is not a shared word. (0)
Yes, the shared word in Spanish is a dedo. (0)
Yes, the shared word in Spanish is adoras. (0)
Yes, the shared word in Spanish is adora. (1)

FIGURE 4.3 A SELECTION OF FOUR ITEMS. ABOVE EACH ITEM IS THE NAME OF THE PROPERTY IT IS USED FOR MEASURING. THE BOLDED LINE FROM THE RESPONSE OPTIONS IS THE CORRECT RESPONSE FOR EACH ITEM

In step 0 of the steps listed above, via Equation 7, these items showed evidence of differential item functioning. The SUM was structured to have four linked test forms. These items all appeared on the same form and the form with the largest sample. In total, 9 items were identified as showing evidence of possible differential item functioning.

Step 1: *Model using an exploratory approach to mixture modeling of all groups combined.* Though, mixture modeling may not be necessary (for instance, one could use other clustering methods), in our case, since step 0 involved a causal latent variable model, so too should this step. The model selection process is a mix of statistical consideration, an idea that models and data should have some fit and plausibility of particular theories. The idea, essentially, is that there me be “non-manifest” DIF and using latent class models that are relatively light on assumptions may be a fruitful method of exploratory enquiry. However, in this case, they should be treated as such. Though, we may treat, for instance, a latent class model as modeling one property that differs in kind (Borsboom et al., 2016), it is also possible to interpret multiple latent classes as corresponding to multiple properties themselves depending on the reason attributed for their being a mixture distribution. This latter interpretation is what is intended here. For instance, among items showing evidence of DIF, we would expect to see multiple latent classes since there are, in effect, two different distributions of item response tendencies for certain items (Ayala et al., 2011 call the grouping variable of interest in a non-latent class context a manifest group).In our conceptualization, the latent classes may reflect the presence of influence properties.

With the full sample of students, a series of latent class mixture models were specified. Mixture models (c.f. Masyn, 2013; McCutcheon, 1987) are a special case of latent variable models in which the specified predictors are categorical in nature and modeled as latent. That is, the properties are treated as latent in the model and predict item responses. In the case of categorical item response data, this subset of models is typically called Latent Class Analysis (LCA). The idea is roughly that there are a subset of classes to be modeled. This is highly exploratory in the sense that one does not pre-specify anything about the classes such as class-specific response probabilities, only the number of classes (though, there are a subset of models that are indeed more confirmatory in nature such as when one might know what the classes might be and we are trying to differentiate or assign subjects to classes). In other words, when using mixture models, the claim is that observed data come from a mixture of distributions each with a given proportion in the population.

To specify, in a mixture model, in the case of our reading example, if we are to specify two classes, the probability of a given correct response to an item, following Masyn (2013), is:

$$p(x_i = 1) \ = [p(c = 1) * p(x_1 = 1|c = 1)] + [p(c = 2) * p(x_2 = 1|c = 2)] + [p(c = k) * p(x_i = 1|c = k)]$$

EQUATION 8

In other words, the probability of answering item *i* correctly is a summation over products of the probability of a student being in class c, effectively, a class size variable and the probability of getting an item right given that a student is in class k. The term $p(c = k)$ is often referred to as the mixing proportion. One can see this is not at the level of the individual student probability. The model can be written in more formal terms for each student *s*, given mixing proportion π of class *k* (summing over π_k yields a probability of 1) and item response *x* (scored 0 or 1) to item *i* by student *s*

$$P(x_{1s}, x_{2s}, \dots x_{is}) = \sum_{k=1}^{K} \pi_k * P(x_{1s}\, , x_{2s} \dots x_{is}|c_s = k)$$

EQUATION 9

The classes are “unknown” prior to the model estimation. There is an assumption that items, conditional on class, are uncorrelated. In this specification, there is no assumption of class ordering, so items can be unordered in the sense that class 1 response probabilities would not have any ordering relation with class 2 response probabilities.

In this step, as mentioned before a series of models are specified each with a differing number of latent classes. Class enumeration proceeded through typical fit procedures and cutoff values with emphasis on the Bayesian Information Criteria (BIC; see for example, Nylund et al., 2007; Schwarz, 1978) and some consideration of the Bootstrap Likelihood Ratio Test (BLRT). The BLRT is a bootstrap method to estimate differences and provide a p-value in the likelihood ratios between enumerating a k-class model and k-1 class model. The BIC is defined as

$$BIC = -2lL + p * log(n)$$

In the BIC equation, lL is the log-likelihood, p is the number of (free) parameters in the model, and n is the number of subjects (or sample size).

Substantively, here, we have some idea that something associated with whether students speak Spanish at home or not influences response probabilities. However, we have not actually specified the particular item features or specific cognitive process that causes this (though, this is a bit easier in this case, as opposed to just a demographic category such as “white”). Tukey (1977) emphasized graphical methods as a means of examining data. Here, we place it in the context of generating hypotheses from the model-based graphical displays of item responses. Here, we are moving between confirmatory methods (Rasch modeling, DIF detection) to more exploratory methods less dependent on model testing (though, model comparison is still occurring). Besides technical class selection criteria, a goal then will be from the perspective of fruitfulness – will particular correct model selection (class selection) allow us to make coherent inferences? Here, part of the goal is property individuation. A number of classes were specified, though, below, for ease of visualization, only the first 6 classes are specified.

Step 2:

Here, the goal is to rerun step 1 but by subgroup. In our case, the full sample includes all students who took the SUM while the subgroups are students who identified as speaking Spanish or English at home as their primary language. This is three groups. While it may be the case that, in theory, running analyses with the full sample of students should yield the same class solutions as the enumeration procedure in step 1, it may not be the case. It is ideally hoped here that the more specific model applying to a particular sample of students may help us reason between subgroups. Here, we are looking to see if, for instance, in the Spanish (or English) speaking groups, there are few if any classes beyond 1 class.

Step 3:

In this step, a multinomial logistic regression is used in which each latent class is effectively regressed on the group membership indicator. In our case, the reference group 0 was were non-Spanish speakers and focal group coded as 1 were Spanish speakers. In our case, a multinomial logistic regression requires multiple steps. A so-called three-step approach was used in which enumeration is followed by uncertainty due to subject class assignment (for instance, whether a student should be assigned to class 1 or class 2) is incorporated. A larger discussion of the complications related to different approaches for including covariates and class assignment of subjects to classes can be found in Nylund-Gibson and Masyn (2016) and Vermunt Click or tap here to enter text. Adapting nomenclature used by Nylund-Gibson and

Masyn (2016) and the indexes and variables for students, items, and classes above, the regression effectively appears:

$$P(c_s = k|Z_s) = \frac{\exp(\gamma_{0k} + \ \gamma_{1k} Spanish_s)}{\sum_{j=1}^{K} \exp\,(\gamma_{0j} + \ \ \gamma_{sj}\, Spanish_s)}$$

Where *Spanish* is a group membership indicator. For this analysis, the group membership indicator included students who marked that they spoke Spanish at home and those students who marked they spoke another language (a majority marked that they spoke English).

This model can be interpreted via comparison, in effect. So, for instance, using a reference latent class (let's say latent class 3), what is the probability of a Spanish speaker being in latent class 2 instead of latent class 3 (in logits). These latent class membership details will be compared to the latent class item probability plots themselves. For instance, we should expect Spanish speakers to answer the cognate items correctly, so should have a higher probability of being in a class that has a higher probability of getting the cognate items correct. In this case, γ_{0k} is an intercept term and γ_{1k} is a coefficient indicating the additional odds of a Spanish speaker being in class k_i instead of class k_j. This model is specified in Figure 4.2, below:

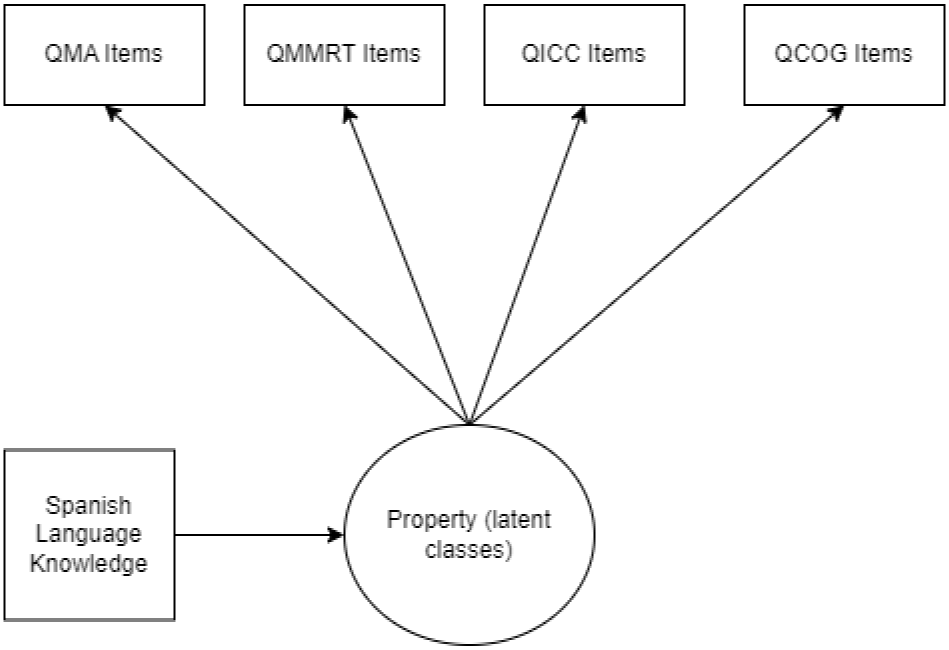


FIGURE 4.4 *THE MEASUREMENT AND STRUCTURAL MODEL SPECIFIED. THE PROPERTY THAT* MAY *BE CAUSALLY RESPONSIBLE FOR ITEM RESPONSES IS THE CIRCLE AND, AT THE VERY LEAST, SPEAKING SPANISH AT HOME IS TAKEN AS SOMEHOW CAUSING OR RELATING TO A VALUE OF THIS PROPERTY (VALUE USED LOOSELY).*

Step 4/5:

Like Thurstone, I do not think classes or clusters should be named. I see this as a starting point for further investigation into the causal relation between items and properties they are meant to elicit. Ideally, in following investigations, items could be written and piloted or interviews performed to start testing some hypotheses generated from the data. Here, we are using a process of inference to the best explanation as posited by Haig (2005). The idea is that the items in this case and the perceived model of item responding limits what we admit into our ontology as causally relevant properties.

This is the graphical comparison step. Here, we plot the response probabilities conditional on the selected models for each group and the full sample. The aim is to identify and reason about different strategies students might use in reading and answering questions from the SUM and how this might differ between groups. If there is more than one class in each grouping and/or the classes are disordered, than student groupings likely should not be treated as homogenous groupings as in a DIF analysis. Since this was exploratory, the initial idea involved looking at items where there were particularly interesting patterns of responding, judged by this researcher. This can help us decipher and more readily articulate the cognitive architecture of responding to certain items (such as what item features might be unrelated to the property of interest). Here, class separation (for instance, are there classes that are far apart in probability for given items?) can be used to identify unique response tendencies. If item response probabilities for a given class are around 50%, this indicates that the response probabilities for that class are relatively uncertain.

Step 6+:

In this step, I will mainly try to reason from modeling efforts which are, to some extent supposed to be representative of the item response process of students, reasons for DIF in these items, and consider what the causes of DIF may be. In other words, we expect to see unique classes since DIF is detected in the items. The influence quantity of Spanish language knowledge is conceived as an influence on the causal relationship between the property and item response. However, it is unclear how this relationship manifests for different properties. For instance, for morphological awareness, whatever properties are associated with speaking Spanish at home may be influence quantities or may be *part* of the attribute of interest. However, it is hard to non-empirically reason about this. For cognate items, one wonders if all Spanish speakers will benefit in the same way from knowing Spanish.

For instance, consider the morphological awareness (MA) item above in Figure 4.1. The made up word is *heterocoloreous*. Thinking from the perspective of section 4.6.4, causal theories of item writing (Mislevy et al., 2017, Chapter 4, called this a design pattern), one can reason that the prefix hetero and root color(eous) can be answered using knowledge of Spanish language in the same way. In this sense, this is just reasoning about the morphemes that does not seem to be apart from the property of morphological awareness. Alternatively, if the emphasis *were* English (though, it is not for the SUM administration), one might consider having a Spanish language vocabulary as a distinguishable property. The following sections will walk through results of the proposed workflow. All analyses were performed in Mplus version 8 (Muthén & Muthén, 2017).

4.6.2 Sample and Instrumentation

Again, for details of the instrument and its piloting see, Arya et al. (2020) . Here, we will work with a subset of 9 items that were identified as showing evidence of DIF (two cognate items, two ICC items, three MA items, and two MMRT items. The SUM had several test forms and the items were selected from the largest form (Total items = 79; total test takers = 1,327). Students were asked what language they spoke at home, and on this form, 328 students indicated speaking Spanish at home as a primary language and 897 indicated speaking another language (of which, 735 indicated speaking English at home and 99 students provided no information).

4.7 Results

I will present results in the order of the steps provided above, skipping step 0 since that is tacitly implied by selection of items. Some steps are part of the analysis as opposed to production of results so will be presented together.

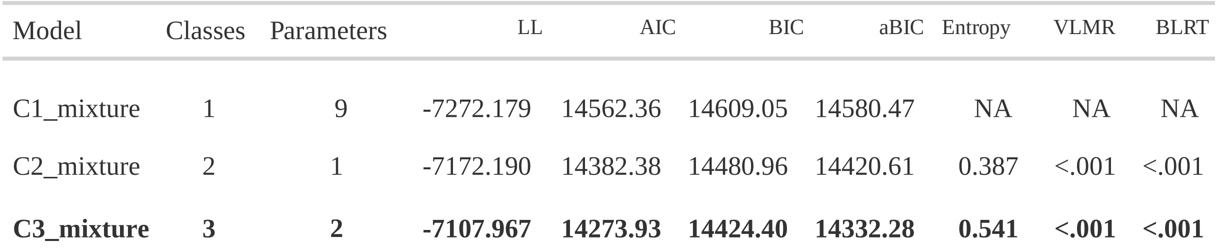
Step 1:

For the full group model, numerous models were specified with varying numbers of latent classes. This includes specifying models with 1 latent class and increasing the number of classes with each iteration. The results are presented in table 4.1, below. In terms of model selection criteria, the distinction between the 3 and 4 class models is relatively slim. However, Figure 4.3, shows profile differences between the 3 and 4 class model are relatively similar, thus, the 3-class model was selected as the final model for simplicity of this illustration.

We would *expect* to see at least one class of students for whom the item response probabilities are higher due to the influence quantity of interest. Here, we would expect there to be students who do indeed have a higher chance of getting the cognate items correct due to Spanish speaking ability and we see that in the green classes in Figure 4.3 for each model.

TABLE 4.1

MODEL FIT INFORMATION FOR MODELS RUNS WITH THE FULL SAMPLE OF STUDENT (N = 1397).



Model	Classes	Parameters	LL	AIC	BIC	aBIC	Entropy	VLMR	BLRT
C1_mixture	1	9	-7272.179	14562.36	14609.05	14580.47	NA	NA	NA
C2_mixture	2	1	-7172.190	14382.38	14480.96	14420.61	0.387	<.001	<.001
C3_mixture	**3**	**2**	**-7107.967**	**14273.93**	**14424.40**	**14332.28**	**0.541**	**<.001**	**<.001**

C4_mixture	**4**	**3**	**-7081.849**	**14241.70**	**14444.05**	**14320.16**	**0.600**	**0.11**	**<.001**
C5_mixture	5	4	-7070.040	14238.08	14492.31	14336.66	0.576	0.51	0.10
C6_mixture	6	5	-7059.866	14237.73	14543.85	14356.43	0.717	0.19	0.27

[1] Note: LL = Log likelihood of the model; AIC = Akaike Information Criteria; BIC = Bayesian Information Criteria; aBIC = Sample Size Adjusted BIC;

VLMR = Vuong-Lo-Mendell-Rubin likelihood ratio rest p-value;

BLRT = Bootstrapped Likelihood Ratio Test P value (does the model with K classes fit better than the model with K-1 classes)

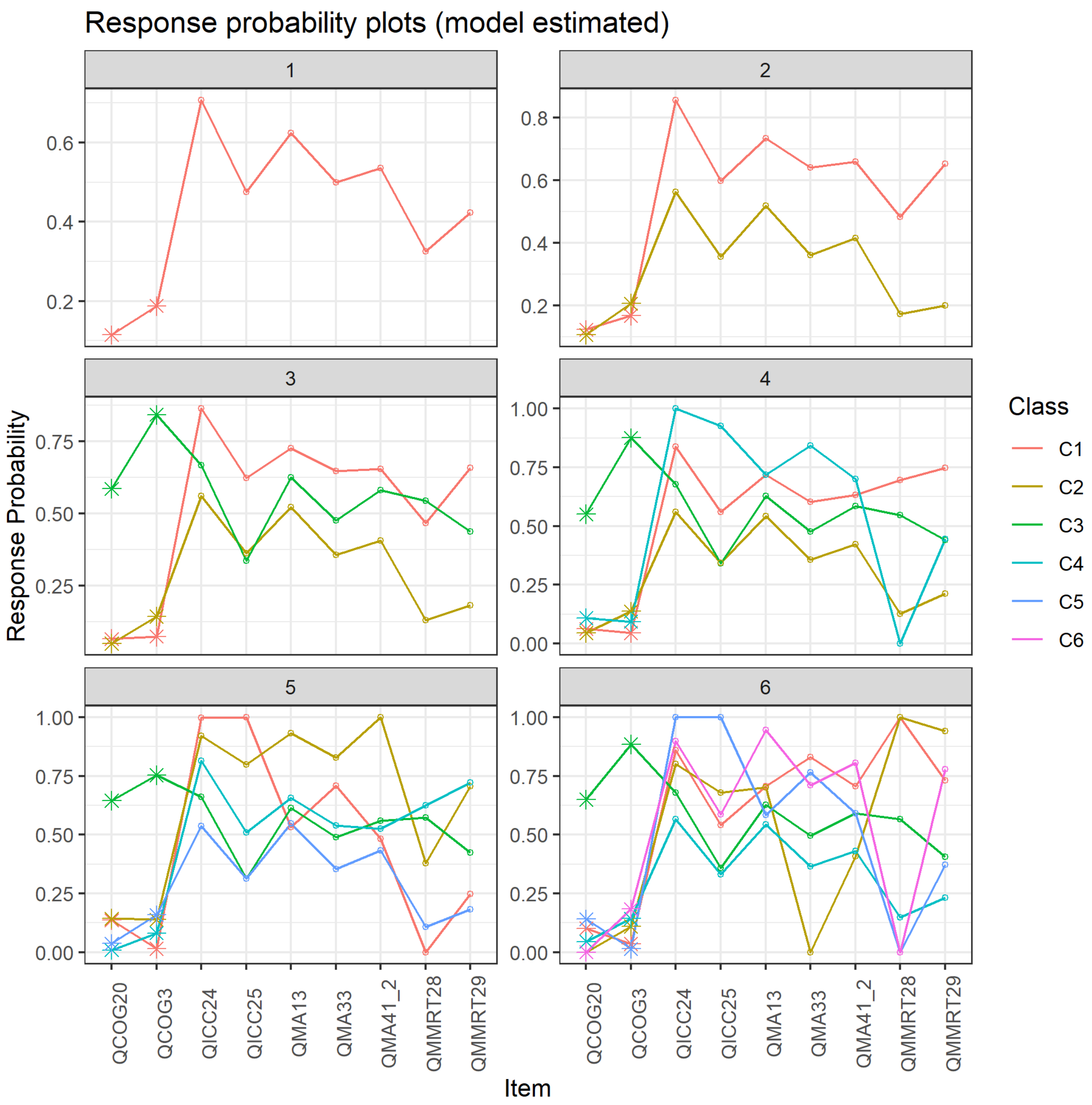


FIGURE 4.5 *RESPONSE PROBABILITY PLOTS FOR ALL MODEL RUNS. THE X-AXIS ARE ITEMS, THE Y-AXIS ARE RESPONSE PROBABILITIES.* EACH PLOT ARE RESPONSE PROBABILITIES AS ESTIMATED BY A SEPARATE MODEL.

Step 2a: Spanish Speakers

In this step, student groups were separated. In effect, the goal of this step is to understand how students might reason about items in different groups. As said above, this is merely meant to be a starting point for generating hypotheses about the response process. Using the concept of DIF which might treat all Spanish speakers as the same, this tries to see if 1) this is a safe assumption, and 2) if it is not, how might we use this to think about the structure of the items and what cognitive processes might we identify as part of the measurand(s).

The table for models with only the students who identify as Spanish speakers at home is presented below, in table 4.2. Again, there is some reason to believe a two-class model (or even a so-called one class model) may be the best fitting model. This would lend some support to the idea that (perhaps unsurprisingly) there may be different cognitive processes and hence different values of the particular property involved that leads to items showing evidence of differential item functioning that may favor students who speak Spanish.

TABLE 4.0

MODEL FIT INFORMATION FOR MODELS RUNS WITH ONLY STUDENTS WHO IDENTIFY AS SPEAKING SPANISH AT HOME (N = 328).

Title	Classes	Parameters	LL	AIC	BIC	aBIC	Entropy	VLMR	BLRT
C1_mixture	1	9	-1937.316	3892.632	3926.769	3898.221	NA	NA	NA
C2_mixture	**2**	**1**	**-1923.814**	**3885.629**	**3957.696**	**3897.429**	**0.711**	**0.60**	**0.02**
C3_mixture	3	2	-1911.682	3881.363	3991.361	3899.373	0.857	0.16	0.23
C4_mixture	4	3	-1902.046	3882.091	4030.019	3906.312	0.691	0.23	0.67
C5_mixture	5	4	-1893.242	3884.484	4070.342	3914.915	0.709	0.20	1.00
C6_mixture	6	5	-1886.504	3891.008	4114.795	3927.649	0.792	0.52	1.00

[1] Note: LL = Log likelihood of the model; AIC = Akaike Information Criteria; BIC = Bayesian Information Criteria; aBIC = Sample Size Adjusted BIC;

VLMR = Vuong-Lo-Mendell-Rubin likelihood ratio rest p-value;

BLRT = Bootstrapped Likelihood Ratio Test P value (does the model with K classes fit better than the model with K-1 classes)

The item probability plots for all model runs are below in Figure 4.4.

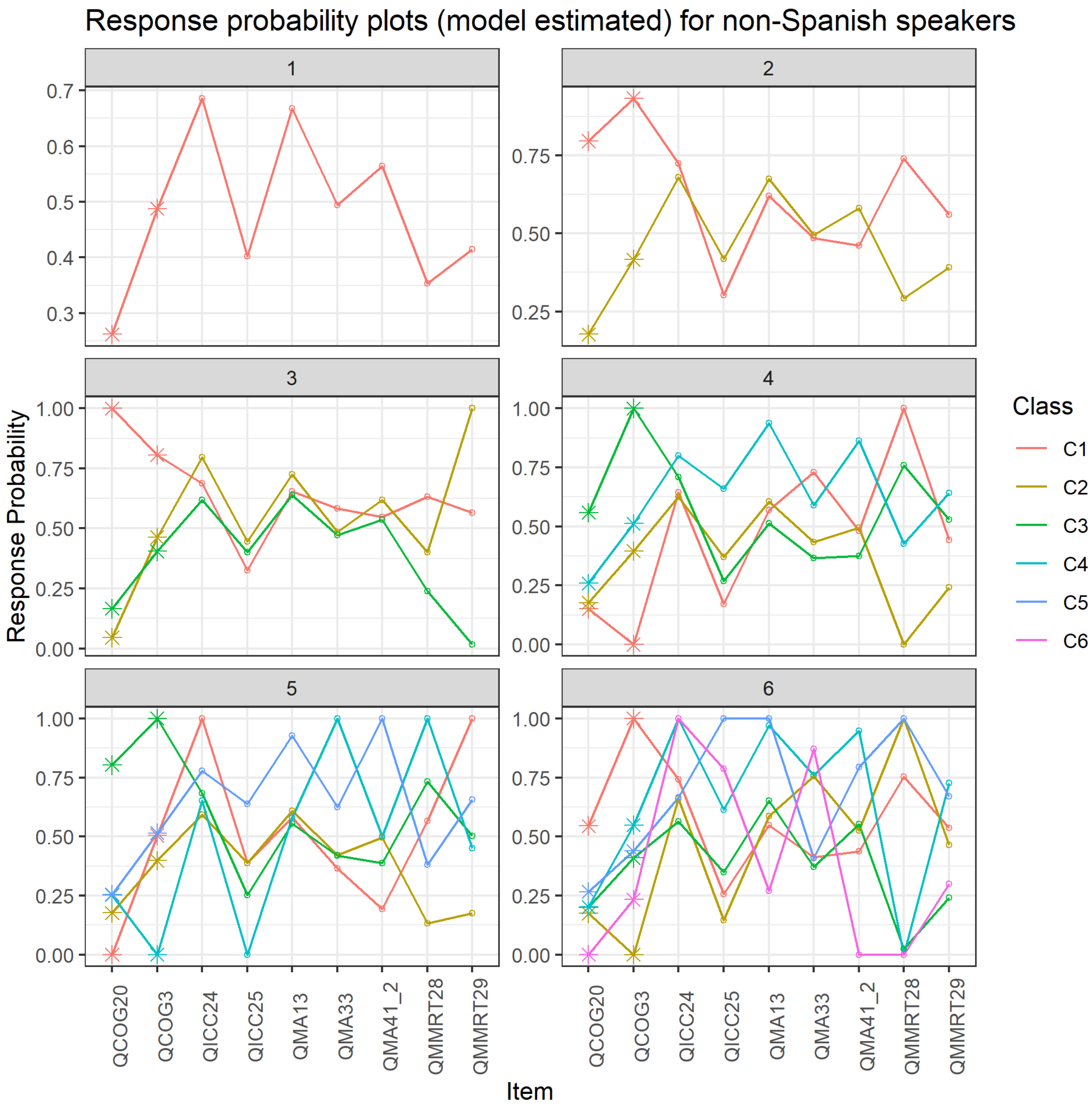


Figure 6.4 Response probability plots for all model runs for just students who identify as speaking Spanish at home. The x-axis are items, the y-axis are model estimated response probabilities. Each plot are response probabilities as estimated by a separate model. The asterisk correspond to the cognate items.

Step 2b: English Language Speakers

The process is the same as step 2a with the same information presented below in table 4.3 and figure 4.5. Here, again, it is hard to differentiate between several models. But since the emphasis is on providing a workflow or proof of concept, a two class model is selected as the best fitting model, though, in the long run has little difference in regards to fruitfulness of later efforts.

Table 4.2

Model fit information for models runs with only students who identify as speaking a language other than Spanish at home (n = 1069).

Title	Classes	Parameters	LL	AIC	BIC	aBIC	Entropy	VLMR	BLRT
C1_mixture	1	9	-4593.763	9205.525	9248.717	9220.134	NA	NA	NA
C2_mixture	**2**	**19**	**-4517.149**	**9072.297**	**9163.479**	**9103.139**	**0.398**	**<.001**	**<.001**
C3_mixture	**3**	**29**	**-4494.484**	**9046.968**	**9186.141**	**9094.042**	**0.633**	**0.27**	**<.001**
C4_mixture	4	39	-4475.375	9028.750	9215.914	9092.056	0.617	0.13	<.001
C5_mixture	5	49	-4463.047	9024.095	9259.249	9103.633	0.580	0.26	0.13
C6_mixture	6	59	-4452.472	9022.943	9306.088	9118.714	0.730	0.46	0.33

[1] Note: LL = Log likelihood of the model; AIC = Akaike Information Criteria; BIC = Bayesian InformationCriteria; aBIC = Sample Size Adjusted BIC; VLMR = Vuong-Lo-Mendell-Rubin-likelihood Ratio Test p value; BLRT = Bootstrapped Likelihood Ratio Test P value (does the model with K classes fit better than the model with K-1 classes

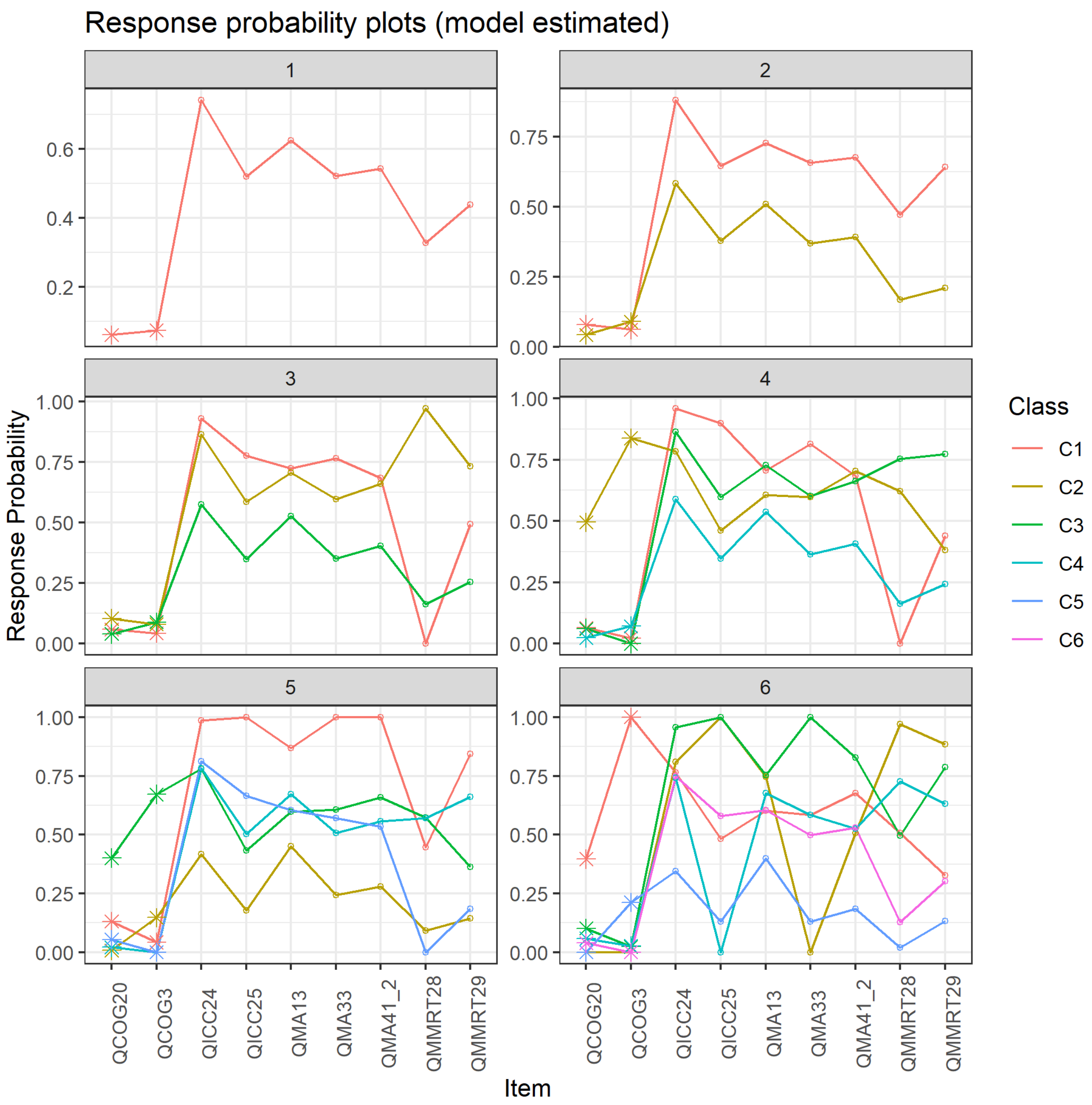


FIGURE 4.7 *RESPONSE PROBABILITY PLOTS FOR ALL MODEL RUNS FOR JUST STUDENTS WHO IDENTIFY AS SPEAKING A LANGUAGE OTHER THAN SPANISH AT HOME. THE X-AXIS ARE ITEMS, THE Y-AXIS ARE MODEL ESTIMATED RESPONSE PROBABILITIES. EACH PLOT ARE RESPONSE PROBABILITIES AS ESTIMATED BY A SEPARATE MODEL.*

Step 3

In this step, using the final model selected with the full sample of students (e.g. the three class model), the latent class variable (in this case, three classes) is regressed on the indicator variable for whether a student spoke Spanish at home or not. That model is visualized in Figure 4.6 below. Here, the reference class is the blue line below. In it, the students have a generally high probability of answering the cognate items. Indeed, the regression results indicate that students are (highly) likely to be in this class. The multinomial logistic regression results are presented in table 4.4.

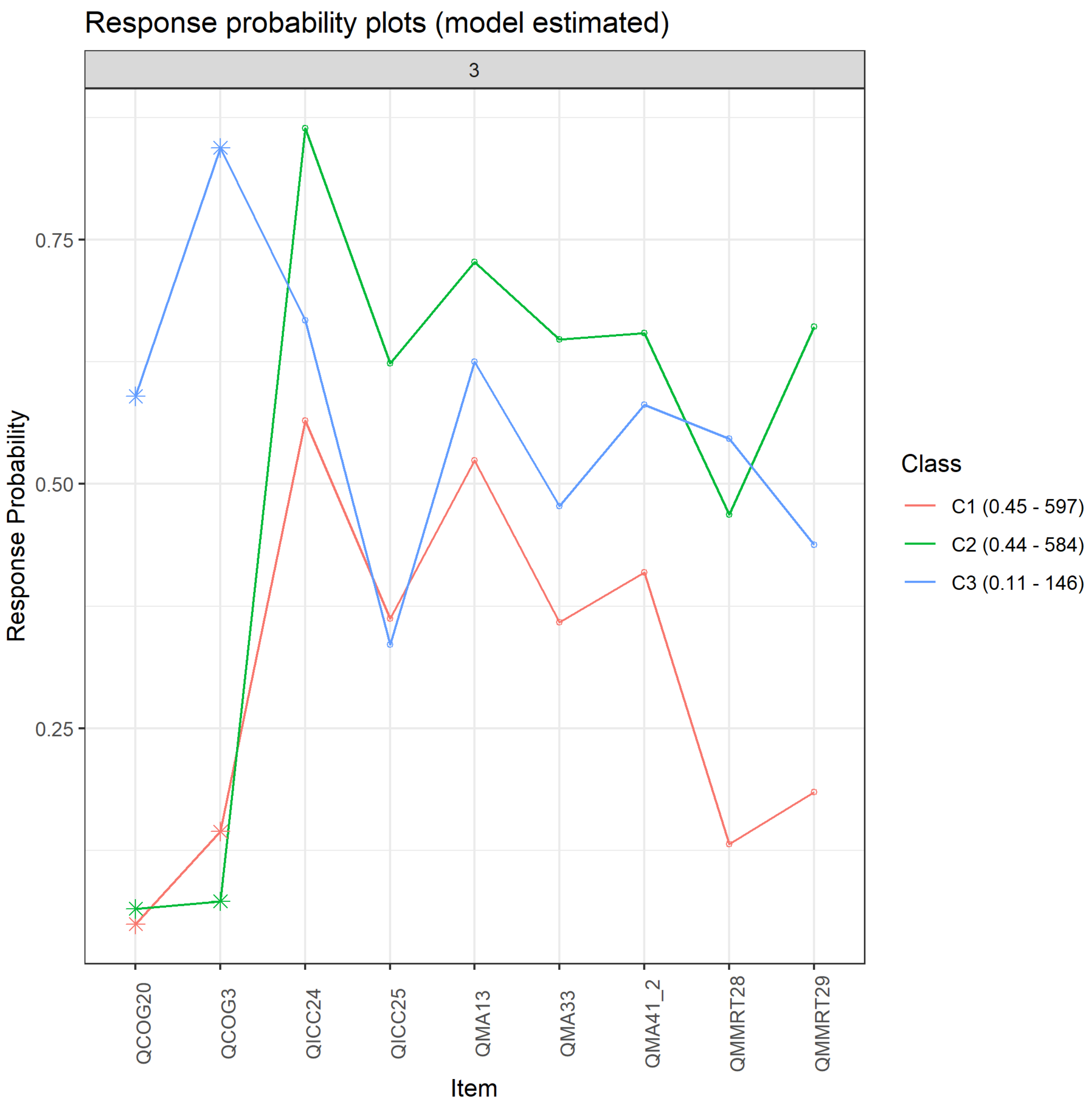


FIGURE 4.8 ***FINAL MODEL SELECTED WITH THE FULL SAMPLE OF STUDENTS. THE LEGEND INDICATES THE CLASS LABEL AND THE PROPORTION OF STUDENTS ESTIMATED TO BE IN EACH CLASS AND THE TOTAL THAT AMOUNTS TO (ROUNDED). THIS CAN BE READ AS CLASS (PROPORTION - TOTAL STUDENTS***

Table 4.4 presents one regression of the probability of each group being in a class relative to class three. Columns 4 and 5 are used to present the probability of Spanish speakers being in class 3 relative to one of the other classes. Indeed as expected, students who identified speaking Spanish at home (Heritage Spanish Speakers) were much more likely to be in Class 3 as opposed to Class 1 or Class 2 (by absurd odds in fact – 28 and 14 times more likely in fact). However, it is worth noting that the class (the green class above) does not have a particularly high probability of answering morphological awareness items correctly (the items that end in MA) compared to the non-heritage Spanish speaker class.

TABLE 4.3

Multinomial logistic regression results in which the outcome variables are the latent classes and the predictor is a dummy coded indicator if a student was a heritage or non-heritage Spanish speaker. The final two column are from the same model, but comparing log odds/odd/probability of Spanish speakers being class 3 relative to a different reference class (in parentheses).

	Reference Class: 3 (High Cognate Probability)		Reference Class: 1	Reference Class: 2
Estimates	Class 1	Class 2	Class 3 (1)	Class 3 (2)
Intercept (Non-heritage Spanish Speakers)				
Estimate (log odds)	2.94	2.82		
Standard Error (log odds)	0.38	0.38		
Test Statistic	7.84	7.47		
odds	18.86	16.76		
Probability	0.95	0.94		
p_value	p < .0001	p < .0001		
Heritage Spanish Speakers				
Estimate (log odds)	-3.32	-2.69	3.32	2.69
Standard Error (log odds)	0.44	0.42		
Test Statistic	-7.58	-6.4		
Odds	0.04	0.07	27.77	14.78
Probability	0.03	0.06	0.97	0.94
p_value	p < .0001	p < .0001	p < .0001	p < .0001

Step 4/5:

I aim to use the graphical comparison step combined with Step 3 to effectively compare final models in the three model runs (model run with all students, just students who claim to speak Spanish at home, and non-Spanish speakers at home). Figure 4.7, below, presents a comparison of all final models. By Final model I mean model that I as the researcher selected. The right columns is, plot B and plot D of Figure 4.7, the same plot to enable comparison of response profiles across Heritage Spanish speakers (B/D) and non-Heritage Spanish speakers (A) as well as across Heritage Spanish speakers (B/D) and the full model (C).

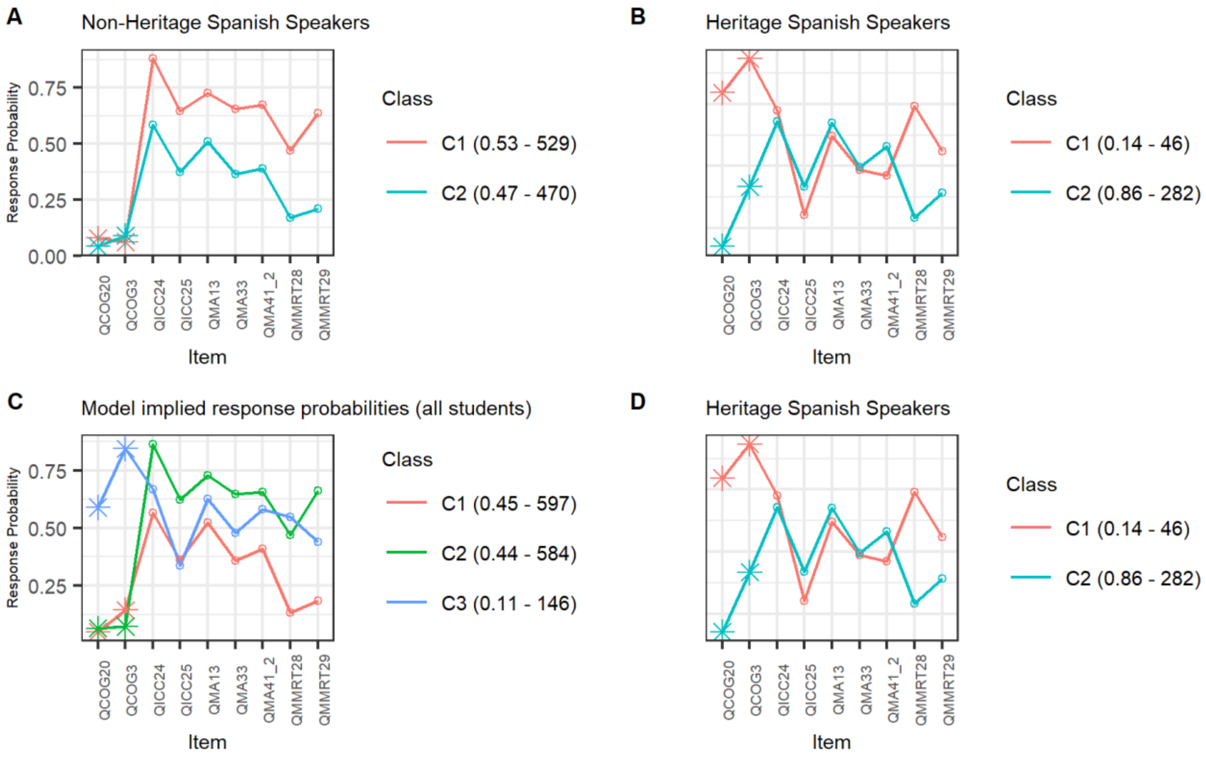


FIGURE 4.9 *RESPONSE PROBABILITIES FOR NON-HERITAGE SPANISH SPEAKERS GIVEN A LATENT CLASS MODEL SPECIFYING TWO CLASSES ARE PRESENTED IN PANEL A. RESPONSE PROBABILITIES FOR A 2 CLASS MODEL FOR SPANISH SPEAKERS ARE PRESENTED IN PANEL B (AND D; PRESENTED TWICE FOR EASE OF COMPARISON). RESPONSE PROBABILITIES FOR A 3 CLASS MODEL FOR ALL STUDENTS ARE PRESENTED IN PANEL C.*

It is worth noting that in these models, there are no prior assumptions about who might be someone who identifies as a Spanish speaker and what their response profile looks like. It is worth remembering that 328 students identified as Spanish speakers at home. Based on Step 3, Spanish speakers were most likely to be in class labeled C3 (blue) in Plot C (bottom right) of Figure 4.7. These are students who were most likely, compared to the other classes, to get the Cognate items correct (which ask for shared meanings of Spanish and English words) – see figure 4.1. Note, however, that approximately 328 students claimed to speak Spanish at home while in plot C, we can see that only 11% or 146 students are classified as being in class C3. And, in the model involving just Spanish speakers (Figure 4.7 B/D) only a small number of students (of course, the sample size is quite small here) were predicted to be in the similarly shaped response profile in the full sample (14% of those who identified as Spanish speakers). This group is also present in the three class model. Of course, the stability of these models is questionable given so many model runs. Nonetheless, we can point an eye towards item structure. We can argue, indeed, that there is an influence quantity involved in this measurement process as indicated by panel C of figure 4.7.

4.8 Discussion (Step 6+)

What might normally go in a discussion section of an empirical paper is taken to be a step

6 (or more) in our example (this analysis may be part of an abductive process). We could term this the inference to the best explanation step (Haig, 2005, 2005). While we may term this step "phenomena detection," it is important to see this as hypothesis generation about the measurand as opposed to hypothesis testing about the measurand. Turning back to step 5 above we can interpret the plots from a causal perspective. That is, there seems to be a group of students, even among Spanish speakers but also non-Spanish speakers who are more likely to get cognate items right and this is caused by some unique property or property value (namely, the class Spanish speakers were most likely to be a part of). Otherwise, response profiles are pretty similar. Perhaps this "property" of cognate knowledge may be no different than the primary causal property of interest related to heritage speakers of Spanish that we might posit is the causal property causing DIF in other items. Saying that, we see that any DIF effect that is large may be caused by whatever *might* be represented by the blue line (C3) in plot 4.7C.

Realistically, though, this is perhaps a job of reasoning by committee. Instead of somebody (e.g. me) looking at these plots and tilting their heads, a discussion needs to occur to consider possible hypotheses about the causes of the plotted reading profiles above. One view here is to consider how we interpret the DIF equation in equation 7 if the heritage Spanish speaker indicator is interacted with each item. In that model, the Spanish language grouping variable is treated as a separate effect from the property of interest. However, reasoning from the data and modeling, we might surmise that something related to speaking Spanish at home is indeed perhaps the same property involved in making sense of item morphemes even if those morphemes have some latinate roots. In that case, it may not even be appropriate to remove items that show evidence of DIF.

Consider the item that involves the made-up word *heterocoloreous*, which is QMA13 in the plots above. There seems to be no real relation with the class of students who have the highest probability of answering cognate items correctly and answering this question correctly even though it involves roots that could be found in English and Spanish (implying the DIF effect may be quite small). In this case, we might hypothesize that Spanish language knowledge and English language knowledge plus bilingualism *may* be part of the measurand since a student might use it in the same way as a non-heritage Spanish speaker. This is based on reasoning about response profiles in the cognate items, effectively using data from one section of the test to make inferences about others.

For instance, consider the cognate item QCOG20 asks:

> Which English word is most similar in meaning to the Spanish word below?
>
> campo

Campo here means *field* though an option is camp, or camper. This might be close to a false cognate, or alternatively, there may be implications about the word campo that are similar to camp. A next step would be to try to refine the measurand by investigating the reason some students, even among non-Spanish speakers apparently (though, of course, there are likely students who mis-marked the question about whether they spoke Spanish at home or the question did not align with their experience), get this item wrong. Though it may ultimately result in changing or removing an item, it would take the form of phenomena detection trying to generally answer the question – “are influence quantities that are distinct from the measurand of worry to the extent results are altered?”. The perspective here is not about advantaging or disadvantaging certain groups as people typically see DIF. Since the SUM is not meant to be a high stakes test, the primary motivation would be to provide information to a teacher, for instance, important information they could leverage. So, reasoning about the measurand, we might actually be willing to include some level, or some amount of Spanish language knowledge (as intended by the SUM).

It is also worth noting, that from this analysis, it appears that treating students who speak Spanish language at home as a homogenous class is likely not a safe assumption. Because of this, we should take seriously the consideration of multiple influence properties involved. The structure of the measurands of the SUM instrument are likely conglomerates or composites at a higher level anyway (as opposed to some atomic conception). Given that we cannot treat this group as a homogenous class of students, what should we do? One theory, is in effect, that this is a matter of bilingualism being a complicated, non-continuous and disordered phenomenon. It is perhaps that some students are more adept with certain Spanish language vocabulary correspondence to English vocabulary, since the two cognate items are false cognates of sorts (and their meaning in English as well), that might be of important use. Alternatively, it could be the case that modeling the influence properties as classes is incorrect, or at least, not the place to *finish* hypothesis generation. We could take another step and treat the potential influence properties as continuous. Here, we might model the location and scale of influence property distributions via something like Hedecker's location scale model where the distributions of students in a given group are modeled with distinct means and variances (Hedeker et al., 2008).

.

4.9 Next steps

To move from hypothesis generation about the properties of interest to testing these hypotheses, one route would be to set up studies specifically focused on matters related to bilingualism in different places. Further, indeed, in the context of the case study, one would have to gather linguists and literacy experts to better unpack the definition of the measurand. These informants would have to be walked through the meaning of the plots produced above to surmise what properties might be responsible for the patterns and how these properties are structured (if any) – for instance, might one of these properties not be a unique property at all but just differing levels of Spanish language fluency or knowledge.

In terms of epistemic and methodological iteration, we might articulate the above effort as a new step in starting to understand how students interact with an instrument and then trying to reason or begin to generate hypotheses from this step about what this phenomenon might be. Of course, we aimed to use some sort of abductive process of reasoning, though, as is a typical criticism of inference to the best explanation – the best explanation may still not be a very good one. This is a modeling effort and not so much a measurement effort. Instead, a more purposeful effort might be necessary. This purposeful effort might involve the development of items in so-called experimental DIF studies in which items are specially developed to test DIF theories. Here, we might be able to add more items that involve roots that share increasing numbers, for instance, of Spanish or Latinate roots vs those that do not.

Finally, there are a number of statistical considerations above. For instance, considering whether, indeed, the methods above are coherent in mixture modeling would require far more work such as simulation related to how classes are recovered in differing scenarios where relative or non-relative measurement might be possible. Conceptually at least, it is hoped that the discussion above provide some notion of how to think about how we might target the definition of an attribute of interest via empirical means. Here we were motivated by notions of exploratory data analysis, influence quantities, and epistemic and methodological iteration. However, I also suspect that there is a large moral or ethical element. Whether one should consider something as part of a measurand definition in the social sciences will be reliant on a model that allows us to see the world in a certain way. Who makes this model, of course, will dictate whether something is considered a nuisance in one regard, or a strength in another.

References

Aekerman, T. A. (1992). A Didactic Explanation of Item Bias, Item Impact, and Item Validity From a Multidimensional Perspective. In *Journal of Educational Measurement Spring* (Vol. 29, Issue 1).

Afflerbach, P., Pearson, P. D., & Paris, S. G. (2008). Clarifying Differences Between Reading Skills and Reading Strategies. *The Reading Teacher*, *61*(5), 364–373. https://doi.org/10.1598/RT.61.5.1

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? In *Medical care* (Vol. 42, Issue 1 Suppl). https://doi.org/10.1097/01.mlr.0000103528.48582.7c

Angoff, W. H. (1993). *Perspectives on differential item functioning methodology.*

Arya, D., Clairmont, A., Katz, D., & Maul, A. (2020). Measuring Reading Strategy Use. *Https://Doi-Org.Proxy.Library.Ucsb.Edu:9443/10.1080/10627197.2019.1702464*, *25*(1), 5–30. https://doi.org/10.1080/10627197.2019.1702464

Ayala, R. J. de, Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2011). Differential Item Functioning: A Mixture Distribution Conceptualization. *Http://Dx.Doi.Org.Proxy.Library.Ucsb.Edu:2048/10.1080/15305058.2002.9669495*, *2*(3–4), 243–276. https://doi.org/10.1080/15305058.2002.9669495

Bollen, K. a. (2002). Latent Variables in Psychology and the Social Sciences. *Annu Rev Psychology*. https://doi.org/10.1146/annurev.psych.53.100901.135239

Borsboom, D. (2005). Measuring the Mind. In *Measuring the mind: Conceptual issues in contemporary psychometrics*. https://doi.org/10.1017/CBO9780511490026

Borsboom, D. (2008). Latent Variable Theory. *Measurement*, *6*, 25–53. https://doi.org/10.1080/15366360802035497

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2002). Different Kinds of DIF: A Distinction Between Absolute and Relative Forms of Measurement Invariance and Bias. *Applied Psychological Measurement*, *26*(4), 433–450. https://doi.org/10.1177/014662102237798

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The Theoretical Status of Latent Variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Borsboom, D., Rhemtulla, M., Cramer, A. O. J., van der Maas, H. L. J., Scheffer, M., & Dolan, C. v. (2016). Kinds versus continua: a review of psychometric approaches to uncover the structure of psychiatric constructs. *Psychological Medicine*, *46*(8), 1567–1579. https://doi.org/10.1017/S0033291715001944

Brennan, R. L. (2000). (Mis) conceptions about generalizability theory. *Educational Measurement: Issues and Practice*.

Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, *44*(2), 131–155.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Davies, A. (2010). Test fairness: a response. *Language Testing*, *27*(2), 171–176. https://doi.org/10.1177/0265532209349466

de Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach* (Vol. 10). Springer.

Elliott, K. C. (2012). Epistemic and methodological iteration in scientific research. *Studies in History and Philosophy of Science Part A, 43*(2), 376–382.

French, B. F., & Finch, W. H. (2009). *Confirmatory Factor Analytic Procedures for the Determination of Measurement Invariance*. https://doi.org/10.1207/s15328007sem1303_3

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

Giere, R. N. (2009). An agent-based conception of models and scientific representation. *Synthese 2009 172:2, 172*(2), 269–281. https://doi.org/10.1007/S11229-009-9506-Z

Giordani, A., & Mari, L. (2012). Measurement, models, and uncertainty. *IEEE Transactions on Instrumentation and Measurement*, *61*(8), 2144–2152.

Gouvea, J., & Passmore, C. (2017). 'Models of' versus 'Models for': Toward an Agent-Based Conception of Modeling in the Science Classroom. *Science and Education*, *26*(1–2), 49–63. https://doi.org/10.1007/S11191-017-9884-4/TABLES/1

Haig. (2013). Detecting Psychological Phenomena: Taking Bottom-Up Research Seriously. *The American Journal of Psychology*, *126*(2), 135. https://doi.org/10.5406/amerjpsyc.126.2.0135

Haig, B. D. (2005a). Exploratory Factor Analysis, Theory Generation, and Scientific Method. *Multivariate Behavioral Research*, *40*(3), 303–329. https://doi.org/10.1207/s15327906mbr4003_2

Haig, B. D. (2005b). An abductive theory of scientific method. *Psychological Methods*, *10*(4), 371–388. https://doi.org/10.1037/1082-989X.10.4.371

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An Application of a Mixed-Effects Location Scale Model for Analysis of Ecological Momentary Assessment (EMA) Data. *Biometrics*, *64*(2), 627. https://doi.org/10.1111/J.1541-0420.2007.00924.X

Holland, P. W. (1990). On the sampling theory roundations of item response theory models. *Psychometrika*, *55*(4), 577–601. https://doi.org/10.1007/BF02294609

Hopewell, S. (2011). Leveraging bilingualism to accelerate English reading comprehension. *International Journal of Bilingual Education and Bilingualism*, *14*(5), 603–620.

Joint Committee for Guides in Metrology (JCGM). (2008). *JCGM 100:2008, Evaluation of measurement data—Guide to the expression of uncertainty in measurement (GUM)*. Sèvres: JCGM.

Juel, C., & Minden-Cupp, C. (2000). Learning to Read Words: Linguistic Units and Instructional Strategies. *Reading Research Quarterly*, *35*(4), 458–492. https://doi.org/10.1598/RRQ.35.4.2

Keller, E. F. (2000). Models of and Models for: Theory and Practice in Contemporary Biology. *Philosophy of Science, 67*, S72–S86. http://www.jstor.org/stable/188659

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363.

Klingner, J. K., & Vaughn, S. (1999). Promoting reading comprehension, content learning, and English acquisition though Collaborative Strategic Reading (CSR). *The Reading Teacher*, *52*(7), 738–747.

Kunnan, A. J. (2007). Test Fairness, Test Bias, and DIF. *Language Assessment Quarterly*, *4*(2), 109–112. https://doi.org/10.1080/15434300701375865

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores. In *Statistical theories of mental test scores.* Addison-Wesley.

Lovasz, N., & Slaney, K. L. (2013). What makes a hypothetical construct "hypothetical"? Tracing the origins and uses of the 'hypothetical construct' concept in psychological science. *New Ideas in Psychology*, *31*(1), 22–31. https://doi.org/10.1016/J.NEWIDEAPSYCH.2011.02.005

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory: *Https://Doi-Org.Proxy.Library.Ucsb.Edu:9443/10.1177/00031224211004187*, *86*(3), 532–565. https://doi.org/10.1177/00031224211004187

Maraun, M. D. (1996). Metaphor taken as math: Indeterminancy in the factor analysis model. *Multivariate Behavioral Research*, *31*(4), 517–538.

Maraun, M. D., & Gabriel, S. M. (2013). Illegitimate concept equating in the partial fusion of construct validation theory and latent variable modeling. *New Ideas in Psychology*, *31*(1), 32–42. https://doi.org/10.1016/j.newideapsych.2011.02.006

Mari, L. (2006). On the measurand definition. *18th IMEKO World Congress 2006: Metrology for a Sustainable Development. Vol. 3.*, *3*, 2518–2519.

Mari, L. (2013). A quest for the definition of measurement. *Measurement: Journal of the International Measurement Confederation*, *46*(8), 2889–2895. https://doi.org/10.1016/j.measurement.2013.04.039

Mari, L., Wilson, M., & Maul, A. (2021). *Measurement across the sciences: developing a shared concept system for measurement*. Springer.

Markus, K. A., & Borsboom, D. (2013). Reflective measurement models, behavior domains, and common causes. *New Ideas in Psychology*, *31*(1), 54–64. https://doi.org/10.1016/J.NEWIDEAPSYCH.2011.02.008

Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In *The Oxford handbook of quantitative methods* (Vol. 2). Oxford University Press Oxford.

Maul, A., Mari, L., Torres Irribarra, D., & Wilson, M. (2018). The quality of measurement results in terms of the structural features of the measurement process. *Measurement*, *116*, 611–620. https://doi.org/10.1016/J.MEASUREMENT.2017.08.046

McCutcheon, A. L. (1987). *Latent class analysis* (Issue 64). Sage.

McDonald, R. P. (2003). Behavior domains in theory and in practice. *Alberta Journal of Educational Research*, *49*(3).

McGrane, J. A., & Maul, A. (2020). The human sciences, models and metrological mythology. *Measurement: Journal of the International Measurement Confederation*, *152*. https://doi.org/10.1016/j.measurement.2019.107346

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*(2), 300.

Meredith, W. (1993). Measurement Invariance, Factor Analysis, and Factorial Invariance. In *PSYCHOMETRIKA* (Vol. 58, Issue 4).

Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., Rose, D., Gravel, J., Colker, A. M., Rutstein, D., & Vendlinski, T. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation*, *19*(2–3), 121–140. https://doi.org/10.1080/13803611.2013.767614

Mislevy, R. J., Haertel, G., Riconscente, M., Rutstein, D. W., & Ziker, C. (2017). *Assessing model-based reasoning using evidence-centered design: a suite of research-based design patterns*. Springer.

Miyake, T. (2015). Reference models: Using models to turn data into evidence. *Philosophy of Science*, *82*(5), 822–832. https://doi.org/10.1086/683322

Muthén, L. K., & Muthén, B. O. (2017). *Mplus: Statistical Analysis with Latent Variables: User's Guide (Version 8). Los Angeles, CA: Authors*.

National Assessment Governing Board. (2021). Reading Framework for the 2026 National Assessment of Educational Progress. In *NAEP*.

Nisbet, I. (2019). Fairness takes centre stage. *Assessment in Education: Principles, Policy & Practice*, *26*(1), 111–117. https://doi.org/10.1080/0969594X.2017.1358151

Nisbet, I., & Shaw, S. D. (2019). Fair assessment viewed through the lenses of measurement theory. *Assessment in Education: Principles, Policy & Practice*, 1–18. https://doi.org/10.1080/0969594X.2019.1586643

Nitko, A. J. (1995). Curriculum-based continuous assessment: a framework for concepts, procedures and policy. *Assessment in Education*, *2*(3), 321–337.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. https://doi.org/10.1080/10705510701575396

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 782–797.

Orilia, F., & Paolini Paoletti, M. (2022). Properties. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022). Metaphysics Research Lab, Stanford University.

Paek, I., & Wilson, M. (2011a). Formulating the Rasch Differential Item Functioning Model Under the Marginal Maximum Likelihood Estimation Context and Its Comparison With Mantel–Haenszel Procedure in Short Test and Small Sample Conditions. *Educational and Psychological Measurement*, *71*(6), 1023–1046. https://doi.org/10.1177/0013164411400734

Paek, I., & Wilson, M. (2011b). Formulating the Rasch Differential Item Functioning Model Under the Marginal Maximum Likelihood Estimation Context and Its Comparison With Mantel–Haenszel Procedure in Short Test and Small Sample Conditions. *Educational and Psychological Measurement*, *71*(6), 1023–1046. https://doi.org/10.1177/0013164411400734

Perfetti, C., & Stafura, J. (2014). Word Knowledge in a Theory of Reading Comprehension. *Scientific Studies of Reading*, *18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

Poehner, M. E. (2011). Dynamic Assessment: fairness through the prism of mediation. *Assessment in Education: Principles, Policy & Practice*, *18*(2), 99–112. https://doi.org/10.1080/0969594X.2011.567090

Ramirez, J. D. (2000). Bilingualism and literacy: Problem or opportunity? A synthesis of reading research on bilingual students. *Proceedings of A Research Symposium on High Standards in Reading for Students From Diverse Language Groups: Research, Practice & Policy*, *33*.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. The University of Chicago Press.

Rigdon, E. E., Becker, J.-M., & Sarstedt, M. (2019). Factor Indeterminacy as Metrological Uncertainty: Implications for Advancing Psychological Measurement. *Multivariate Behavioral Research*, *54*(3), 429–443. https://doi.org/10.1080/00273171.2018.1535420

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.

Searle, J. R. (1995). *The construction of social reality*. Simon and Schuster.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika 1993 58:2*, *58*(2), 159–194. https://doi.org/10.1007/BF02294572

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies in History and Philosophy of Science Part A*, *42*(4), 509–524. https://doi.org/10.1016/J.SHPSA.2011.07.001

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.

Thurstone, L. L. (1940). Current issues in factor analysis. *Psychological Bulletin*, *37*(4), 189.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA.

van Assche, E., Duyck, W., Hartsuiker, R. J., & Diependaele, K. (2009). Does bilingualism change native-language reading? Cognate effects in a sentence context. *Psychological Science*, *20*(8), 923–927.

van Fraassen, B. C. (2012a). Modeling and measurement: The criterion of empirical grounding. *Philosophy of Science*, *79*(5), 773–784.

van Fraassen, B. C. (2012b). Modeling and measurement: The criterion of empirical grounding. *Philosophy of Science*, *79*(5), 773–784. https://doi.org/10.1086/667847

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450–469.

[VIM3] 2.3 measurand. (n.d.). Retrieved August 5, 2022, from https://jcgm.bipm.org/vim/en/2.3.html

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Lawrence Erlbaum Associates.

Wilson, M. (2013). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement: Journal of the International Measurement Confederation*, *46*(9), 3766–3774. https://doi.org/10.1016/j.measurement.2013.04.005

Zumbo, B. D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832

Zumbo, B. D. (2017). Trending away from routine procedures, toward an Ecologically Informed In Vivo View of Validation Practices. *Measurement: Interdisciplinary Research and Perspectives*, *15*(3–4), 137–139. https://doi.org/10.1080/15366367.2017.1404367

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). *A Methodology for Zumbo's Third Generation DIF Analyses and the Ecology of Item Responding*. https://doi.org/10.1080/15434303.2014.972559

Chapter 5

Conclusion

In this section, I aim to emphasize a few key points from the following four chapters and briefly comment on the nature of statistics and measurement. Namely, I aim to complete the connection between the ethical imperative of paying mind to definitions of that which we would like to measure, empirical statistical work, and notions of psychometrics as purely statistical. The empirical work of psychometrics and the conceptual work of defining what we aim to measure make psychometrics a measurement-oriented endeavor as opposed to just statistics. Some of this conceptual work may be of an ethical variety. The question, <*should we measure*?> should not be a question for others outside of psychometrics as well as psychometricians unless it is made clear that which we would like to measure. I aim to wrap up with a few recommendations.

In chapter 1, I tried to set up the problem of definition in psychometrics. The subject of connections between measurement and science was grazed as was the idea that properties are central to measurement. However, it was shown that some primary texts that instruct psychologists in how to use latent variable models do not give much attention to definitions of attributes. Here, I tried to introduce the two running examples that were present in the following two chapter and set up the case study in Chapter 4 – the definition of reading in

NAEP’s reading test and the definition of resilience, both in different phases of their scientific lifecycle.

In chapter 2, my goal was to set up science as a malleable human-made endeavor that people, not models or machines as central characters. The same could be said of measurement as something that yields privileged information but is indeed a human-oriented process. Perhaps said more strongly, what makes science or measurement in fact science or measurement, is that it is inherently human and pragmatic, serving particular needs. Often, these needs involve defining that which we would like to measure, a demarcation task. However, when studying humans, introspective work and power structures often have a hand in creating the things we study. What is considered important, cultural competences, are dictated by emergent power structures. These power structures can be seemingly innocuous – such as who gets to “create” educational or curricular standards in schools – effectively policy decisions sometimes left to those doing psychometrics. This idea that measurement via assessment has been a power wielding endeavor is perhaps made most obvious by the historical study of age-graded schools via testing (Reese, 2013) to serve a particular purpose. Tests in the United States were constructed to oust a particular group of people and bring in others. In service of the notion of standardization, schools were organized by age. In that sense, schooling in the United States is intrinsically linked to standardized tests by the end of the 19th century.

The ethical strand of measurement faces a cross-roads in psychometrics. If psychometricians take a somewhat rule-following position (Nisbet, 2019; Nisbet & Shaw, 2019), or even legalistic, this need not consider definition of what we aim to measure at all. Others, especially non-psychometricians, may think of fairness of tests from an outcome perspective (consequentialism). We see this in debates how tests should sort students into colleges, for instance, and ultimately jobs. This is ultimately Fricker's (2007) larger point about forms of epistemic injustice. That which is supposedly measured by academic tests, or even the norms of testing, are largely determined by people who reside in educational testing companies, and to some extent, this dictates curricular activities (as would be expected given Reese's 2013, book). In the case of resilience research, a devotion to the structure of Werner's metaphor (1995) may dictate what is a desirable outcome and hence something to study. Were we to look toward standards in education and psychology (AERA, APA, NCME, 2014), we shan't find any guidance for how to define what we would like to measure and do so critically. I say the ethical strand of psychometrics is at a cross roads because of the increasing a-psychological account of psychometrics (Wijsen et al., 2022; Wijsen & Borsboom, 2021) that emphasizes generic statistical models that are not based on a certain modal relevance to the properties we measure. Hence, instead, focusing on statistics and probability need not be a psychological discipline at all.

However, measurement is ultimately about *what* we measure. This was the focus of chapter 3 – which hopefully provided sufficient guidelines for the demarcation task that is definition of properties for measurement. Instead of relying on operationalism, we rely on, in effect, providing enough information that a plurality of measurement methods may interact with the property. Each definitional effort requires its own norms and methods, though, different measurement efforts, given needs for accuracy or something else dictate the specificity of the measurand. I hesitate to provide a checklist. Checklist science may, in fact, not be a science at all. Instead, here, I'd like to focus on general ideas and coherence, that we characterize and identify what we would like to describe as richly as possible, before naming that which we would like to measure (e.g. do not ask "what is resilience?" but perhaps, "what are the causal features that lead to a certain outcome). Paying attention to the language we use might prevent a bewitching or creating of puzzles out of only language (Jost & Gustafson, 1998; Maraun, 1998; Wittgenstein, 2009). For instance, when we debate whether background knowledge should be considered part of the reading comprehension property, this is likely not an empirical debate but a normative one – what ought we to measure. Further, the lack of a concept of the meaning of NAEP results in terms of values of a property of interest (and same with resilience research), the lack of a firm definition of a property of interest, and in some cases, outright contradiction (for instance, mixing up causes with effects), render the debate unproductive. Most importantly perhaps, there is no model of the measurand. In the case of resilience, the debate is purely linguistic, though, there may be hidden empirical questions. The aims of the debate might be better framed in terms of types of definitions. But there is no hidden meaning to be discovered via the empirical world as the meaning of concepts and terms comes from ourselves. Again, here, we see that wanting a final firm definition requires someone with the power to do so – a stipulation that removes power from the world, a fixing of worldly phenomena based on linguistic desires. While, words have their referents and by trying to admit all meanings associated with a term into a term (e.g. all the meanings of *resilience)* only confuses. That is, the linguistic effort is important for understanding a scientist's aims. However, unless one is interested in linguistic practice, for the scientist, the effort of defining should be a utilitarian one – to demarcate something in the world instead of creating that thing. Definition can be a modeling task – definitions in natural language are merely one type of model. This is perhaps a final point to be made on this topic – the rich description should offer a way to model the structure of a property.

In that sense, we can think of any definition, even an operationalist one (defining what we measure by the methods used to measure it), as a model. Aiming to think about a model of measurement and the place of the property that is measured help us concoct useful definitions that are in fact, themselves, fallible, since, for instance, a statistical model should have some resemblance to the model of the measurand (the property we aim to measure) and can help test certain theories about the measurand. That is, we aim to define properties so that we can measure, model, and make claims about the property beyond the scope of instances of scientific investigation. The very statistical models we use to investigate phenomena, then, are dictated by the definitions of the measurand. Since models are themselves abstractions, it means they can always be improved, as can the definition of the measurand. Hopefully, one can see how this was the case in chapter 4 case study. The empirical example and ideas about how to model, even when exploratory, were dictated by definitions of what was measured by a particular measuring instrument. Most importantly, by having a definition that was acknowledged for its imperfections, or at least, for the fact that it was an abstraction of some sort, shall hopefully allow for epistemic and methodological iterations. While I used one example with one particular instance of reading, this need not be how epistemic iteration proceeds. In fact, one could have accomplished something perhaps more coherent by speaking with students directly about meanings of words across multiple languages. Of course, whether the mixture modeling approach is merely a re-use of data that identifies certain regularities but not necessarily actual properties could be a fair critique. All the same, it is hoped that it can be seen that moving from a measurement scenario to using the data for more of a theoretical modeling role might help define the property of interest, or at least be one potential way of showing that definitional work, while conceptual, can still have empirical elements. Of course, empirical elements need not be limited to quantitative methods. Nonetheless, the *interpretation* of the results via inference to the best explanation is ultimately a matter of conceptual work that is very much part of the analysis.

This returns to the question of whether psychometrics is about statistics or psychology. If measurement and properties are inherently linked, then there is certainly no way that psychometric models found in journals like *Psychometrika* that are derived not from experience of psychological properties but from rules of probability, can be said to be about measurement. We can see then that the psychometrician may rid themselves of substantive concerns and mistake their/our statistical models as substantive models of the things we measure. It is perhaps of the utmost importance in psychometrics that we see these models absent any substantive input, as not psychometric models, but just, statistical models. This is not to say statistical work is unimportant, in fact it's been quite important and fruitful. Instead, this statement is primarily meant to limit the authority we give to findings from applying off-the-shelf statistical models, encouraging the maintenance of an ever-critical eye that does not accept the voracity of these models in any setting because it can you over some non-epistemic hurdles (again, see chapter 3 but also, Thurstone, 1940).

Of course, part of the challenge may indeed be that data is often taken as "raw" when in fact, even data itself is model-based. That is, data is collected always with a certain aim – to find out about particular elements of the world (Miyake, 2015)1. It may seem that the model only becomes present after the fact, or after the data is collected. However, the reason the statistical model is applicable at all is because of an initial construction of ideas about how data could become evidence of some phenomenon. This is the motivation of, for instance, construct mapping (Wilson, 2005), which is, in effect, a cognitive models of item responding in which relevant observations are sought out based on the model's dictation of what's *relevant.* This is not to say, even in this world, that exploration is not possible. The difference is that most of the time, we are not fossil hunters excavating and being surprised by what we find buried in the data giving us a murky tale. The data is, instead, itself constructed, typically, in the human sciences, in a way that is quite intentional. In future work, I hope to differentiate between modeling a measurand and its structure, modeling the structure of a measurement operation or process, and make further connection to these two points in terms of semiotics and meaning in the context of human measurement. Further, I would primarily like to do this in the context of actual measurement work in the human sciences.

An interesting guiding directive comes from Nguyen's paper (2019) on the philosophy of games which differentiates between achievement play and striving play. It is perhaps analogous to the games we play in social science and academia. The idea of validation in measurement or fairness research in the AERA, NCME, APA *Standards* can be hampered when we aim for achievement play. Here, achievement play is merely trying to win for the sake of it – or in our case, worrying about meeting criteria that serve as justifications for publication. Instead, perhaps, we should aim for striving play that should encourage epistemic iteration – an ethos that the research is about continually getting better. Nguyen writes:

> "When we engage in aesthetic striving play, we are taking on temporary ends for the sake of the intrinsic value of the experience of struggling … the aesthetic account shows how striving play might be accorded a significant place in a meaningful human

> life, through its capacity to sculpt a unique kind of aesthetic experience" (Nguyen, 2019, p. 7).

While achievement play leaves us little epistemic agency in our own scientific practice because the rules are finite and we are dictated by what gets us a win (publishing, getting to say some measuring device is fair and fit for use), striving play is focused on giving agency back to the scientist (or game player) – dictating what it means to become successful on one's own terms. In this sense, the ideas of science are about continual improvement of the scientist, the game, and the scientist's community. From the perspective of publishing measurement papers, this likely means giving credit to a wider array of work, including phenomenological work, as *measurement* work and accomplishments in their own right.

But how do we know if we've improved? Unfortunately, we may not always. Chang's operational coherence may provide a partial answer (Chang, 2017; el Mawas, 2021), at least. That is, we return to pragmatic realism. Does what we are working on allow us to take action that improves what we can do in the world, continually, across many contexts, with that instrument created, study conducted, or hypothesis refuted?

References

American Educational Research Association, American Psychological Association, & and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Chang, H. (2017). VI—Operational coherence as the source of truth. *Proceedings of the Aristotelian Society*, *117*(2), 103–122. https://doi.org/10.1093/arisoc/aox004

el Mawas, O. (2021). Probing 'operational coherence' in Hasok Chang's pragmatic realism. *European Journal for Philosophy of Science 2021 11:4, 11*(4), 1–29. https://doi.org/10.1007/S13194-021-00425-X

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

Jost, J. T., & Gustafson, D. F. (1998). Wittgenstein's Problem and the Methods of Psychology. *Theory & Psychology*, *8*(4), 463–479. https://doi.org/10.1177/0959354398084002

Maraun, M. D. (1998). The Nexus Misconceived: Wittgenstein Made Silly. *Theory & Psychology*, *8*(4), 489–501. https://doi.org/10.1177/0959354398084004

Miyake, T. (2015). Reference models: Using models to turn data into evidence. *Philosophy of Science*, *82*(5), 822–832. https://doi.org/10.1086/683322

Nguyen, C. T. (2019). Games and the art of agency. *Philosophical Review*, *128*(4), 423–462.

Nisbet, I. (2019). Fairness takes centre stage. *Assessment in Education: Principles, Policy & Practice*, *26*(1), 111–117. https://doi.org/10.1080/0969594X.2017.1358151

Nisbet, I., & Shaw, S. D. (2019). Fair assessment viewed through the lenses of measurement theory. *Assessment in Education: Principles, Policy & Practice*, 1–18. https://doi.org/10.1080/0969594X.2019.1586643

Reese, W. J. (2013). Testing wars in the public schools : a forgotten history . In *Testing wars in the public schools : a forgotten history*. Harvard University Press.

Thurstone, L. L. (1940). Current issues in factor analysis. *Psychological Bulletin*, *37*(4), 189.

Werner, E. E. (1995). Resilience in Development. *Current Directions in Psychological Science*, *4*(3), 81–85. http://www.jstor.org/stable/20182335

Wijsen, L. D., & Borsboom, D. (2021). Perspectives on Psychometrics Interviews with 20 Past Psychometric Society Presidents. *Psychometrika 2021 86:1*, *86*(1), 327–343. https://doi.org/10.1007/S11336-021-09752-7

Wijsen, L. D., Borsboom, D., & Alexandrova, A. (2022). Values in psychometrics. *Perspectives on Psychological Science*, *17*(3), 788–804.

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Lawrence Erlbaum Associates.

Wittgenstein, L. (2009). *Major Works: Lugwig Wittgensteing* (First). Harper Collins Publishers.

Appendix

All code can be found at the link: https://osf.io/se4t5/