Psychological Construct Validity

Caroline Marie Stone

Washington University in St. Louis
Psychological Construct Validity
by
Caroline Stone

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2021
St. Louis, Missouri
Table of Contents

List of Figures .................................................................................................................. v
List of Tables .................................................................................................................... vi
Abstract ........................................................................................................................... viii
Foreword .......................................................................................................................... 1
  0.1 Introduction ............................................................................................................. 1
    0.1.1 Mathematical Theories of Measurement ....................................................... 3
    0.1.2 Measurement theory and psychological measurement .............................. 4
    0.1.3 Psychological Measurement ....................................................................... 8
    0.1.4 Terminology ............................................................................................. 12
Chapter 1: The Coordination Problem ........................................................................... 15
  1.1 Introduction ....................................................................................................... 15
  1.2 Galton’s Intelligence Battery .......................................................................... 19
    1.2.1 Formalizing the problem ...................................................................... 24
  1.3 Data-technique circles at large ....................................................................... 28
    1.3.1 The Experimenters’ Regress ................................................................ 29
    1.3.2 The Importance of Robustness ............................................................... 31
    1.3.3 The Extrapolation Worry ..................................................................... 40
  1.4 Conclusion ....................................................................................................... 51
Chapter 2: Psychological Construct Validity ................................................................. 53
  2.1 Introduction ..................................................................................................... 53
  2.2 An Amorphous Concept ................................................................................. 54
    2.2.1 Cronbach and Meehl (1955): “Construct Validity in Psychological Tests” 56
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.2 The Case for Bottom-up Revision</td>
<td>141</td>
</tr>
<tr>
<td>4.4 Conclusion</td>
<td>144</td>
</tr>
<tr>
<td><strong>Chapter 5: Epistemic Iteration</strong></td>
<td>146</td>
</tr>
<tr>
<td>5.1 Introduction</td>
<td>146</td>
</tr>
<tr>
<td>5.2 Coherentism about Measurement</td>
<td>149</td>
</tr>
<tr>
<td>5.2.1 Historical Perspectives on the Coordination Problem</td>
<td>150</td>
</tr>
<tr>
<td>5.2.2 Epistemic Iteration</td>
<td>153</td>
</tr>
<tr>
<td>5.2.3 Epistemic Iteration in Psychology</td>
<td>159</td>
</tr>
<tr>
<td>5.3 Evaluating Epistemic Iteration</td>
<td>163</td>
</tr>
<tr>
<td>5.3.1 De-fanging the Coordination Problem</td>
<td>163</td>
</tr>
<tr>
<td>5.3.2 Accommodating Psychological Constructs</td>
<td>165</td>
</tr>
<tr>
<td>5.4 Coherence and Epistemic Goals</td>
<td>166</td>
</tr>
<tr>
<td>5.4.1 Varieties of Unification</td>
<td>170</td>
</tr>
<tr>
<td>5.4.2 Unification in Psychology</td>
<td>172</td>
</tr>
<tr>
<td>5.4.3 Implicit Memory</td>
<td>174</td>
</tr>
<tr>
<td>5.4.4 Intelligence</td>
<td>180</td>
</tr>
<tr>
<td>5.4.5 Conflicting Epistemic Goals</td>
<td>183</td>
</tr>
<tr>
<td>5.5 Coherence from the Perspective of Psychologists</td>
<td>184</td>
</tr>
<tr>
<td>5.5.1 Systems of Practice</td>
<td>184</td>
</tr>
<tr>
<td>5.5.2 Evaluating Epistemic Goals</td>
<td>187</td>
</tr>
<tr>
<td>5.6 Conclusion</td>
<td>192</td>
</tr>
<tr>
<td><strong>Afterword</strong></td>
<td>195</td>
</tr>
<tr>
<td>5.7 Introduction</td>
<td>195</td>
</tr>
<tr>
<td>5.8 Validity vs Legitimacy</td>
<td>195</td>
</tr>
<tr>
<td>5.9 Coherentism</td>
<td>197</td>
</tr>
<tr>
<td>5.10 Conclusion</td>
<td>199</td>
</tr>
<tr>
<td>References</td>
<td>200</td>
</tr>
</tbody>
</table>
List of Figures

Figure 2.1: Construct Validity ................................................................. 70

Figure 5.1: Relationship between implicit and explicit memory.................. 178
List of Tables

Table 3.1: Psychologists on Operationism ...................................................... 98

Table 4.1: Elements of Scientific Theories ..................................................... 119
Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Prof. Carl Craver for his continuous support of my research, his patience, enthusiasm, and immense knowledge.

My sincere thanks goes to my fellow graduate students, Kate Schimdt, Katie Rapier, Christiane Merritt, and Jordan Shaw. Without your friendship, this would not have been possible.

I am indebted to my friend, Dov Lerman-Sinkoff, who provided a roof over my head, and endless emotional support during the final months of work. I am also indebted to my friend Mike Schmidt, who never tired of explaining technical scientific phenomena, or trouble-shooting latex.

I am incredibly thankful for my partner, Max Viau, who carried me over the finish line and was the cheerleader I didn’t know I needed.

Last but not least, I would like to thank my parents, Tom and Larraine Stone. I am eternally grateful for their encouragement, love, and never wavering belief in my abilities. Thank you for encouraging my curiosity that drove me to philosophy.

Caroline Stone
Dedicated to Lina and Tzoof.
ABSTRACT OF THE DISSERTATION

Psychological Construct Validity

by

Caroline Stone

Doctor of Philosophy in Philosophy

Washington University in St. Louis, 2021

Professor Carl Craver, Chair

A primary concern for any psychological research project is determining how to measure unobservable mental entities, such as ‘implicit memory’, or ‘intelligence’. Psychologists say that a measure has construct validity when they believe that a measurement method measures the construct they intend it to measure, where a construct is any theoretical term that refers to a mental entity. Construct validation, then, is process of justifying one’s belief that a measure has construct validity. My dissertation seeks to answer three related questions, 1. What is construct validity?, 2. What is the best epistemic theory of justification for construct validation?, and 3. How does construct validity relate to the justification of constructs?

There are two problems any theory of construct validation must face that my dissertations solves; problems that can be traced back to Cronbach & Meehl’s (1955) seminal paper, “Construct Validity of Psychological Tests”. First, despite the widespread use of the term, there is a significant ambiguity among psychologists about what exactly construct validity is. Second, the practice of construct validation is premised on an inference that many philosophers think is circular (Tal, 2013). How we establish the validity of a measure depends on what the construct being measured is. However, determining what a construct is, i.e. building the surrounding theory that defines the construct, requires already having measures.
of the construct. In other words, the question “Does this test measure intelligence?”, and the
question, “What is intelligence?” are questions that presuppose answers to one other.

I argue that construct validity—the adequacy of a test as a measure of a particular con-
struct—needs to be distinguished from construct legitimacy—the adequacy of the construct
itself, relative to the theory of which it is a part—. This distinction is necessary because while
both interact, in that increasing one feature can increase the other, both respond to distinct
types of evidence.

In order to solve the coordination problem, I evaluate two epistemic theories already utilized
by psychologists: operationalism and hypothesis-testing. Both operationalism and hypothesis-
testing fail because they do not accurately account for the bidirectional relationship between
measure and construct. Operationalism fails to distinguish between measure and construct
at all, while hypothesis-testing cannot accommodate construct revision driven by measures.

Rejecting operationalism and hypothesis-testing, I turn to a more recent contender from the
philosophy of measurement: coherentism. Coherentism, as formulated by Chang (2004), is
the view that measures and constructs are validated in tandem, through a process of mutual
refinement, or \textit{epistemic iteration}. The coordination problem is solved by shifting focus from
the validity of a measure at a single point in time to validation over time. Further, epistemic
iteration accommodates the bidirectional relationship between construct and measure. It
is not just the case that changes in theory cause changes in measurement, but also that
changes in measurement can cause changes in the theory. The primary limitation of epistemic
iteration is that it relies on a vague notion of scientific progress. A revision is considered to
be progressive as long as it coheres with the epistemic goals of a research field. In order to
avoid an overly permissive account in which any revision would be progressive, I advocate
for adapting a coherentist epistemic theory of justification. I show how one such theory,
Thagard’s (2007) explanatory coherence, can be applied to psychology in order to constrain and evaluate epistemic goals.
0.1 Introduction

Measurement is an essential part of scientific inquiry. When a scientist measures something, be it a physical quantity like mass or length, or more abstract quantities like mental effort or intelligence, they seek to gain knowledge of a system by interacting with system in a particular way. For example, one may measure the length of an object by placing a ruler against the object and observing how the marks on the rule align with the object. The length of the object is determined by assigning a numerical quantity, say four inches, based on the observation. This activity tells us something about the object; in this case, that the object is a certain length.

Bertrand Russell defines measurement as:

"...any method by which a unique and reciprocal correspondence is established between all or some of the magnitudes of a kind and all or some of the number, integral, rational, or real." (1903, p. 176)

Magnitudes are, roughly, quantities or sizes of something. The notion of magnitudes dates back at least to Euclid’s *Elements* (Tal, 2017). In most cases, measurement involves
assigning numbers to magnitudes, in such a way that mathematical relationships between numbers on the measurement scale correspond in a meaningful way to structural relations in the measured system. With scale of length (e.g. inches, or centimeters) for example, the mathematical relation of larger than or smaller than corresponds with the physical property of being longer or shorter. Thus an object that is 8 inches is longer than an object that is 2 inches.

The creation of psychology as a scientific field required researchers to create methods of measuring and quantifying in principle unobservable psychological entities, like mental states, attributes, and cognitive capacities (J. Cattell, 15). The question this dissertation aims to answer is how psychologists create these measures and how they justify their use.

The development and validation of these measures is a difficult task. What is difficult is not merely quantifying the unobservable. It is also difficult to create non-viciously-circular evidence for one’s interpretations of a measure.

Here, I want to preface my work with a brief survey of already existing work in the philosophy of measurement in order to situate my project and explain its philosophical significance.

Tal (2017) differentiates between five types of measurement theories: 1. Mathematical, 2. Operationalist/Conventionalist, 3. Realist, 4. Information theoretic, and 5. Model-based. My focus here is on mathematical theories of measurement. Mathematical theories have dominated modern philosophy of measure and have also yielded conclusions and categorizations systems that are used by scientists today, including psychologists. Therefor, mathematical theories are the most relevant to measurement in psychology.
Measurement theory encompasses a highly heterogeneous body of work spanning from the late 1800’s to today. What connects these disparate theories is the emphasis on the mathematical nature of measurement. Mathematical theories of measurement (referred collectively to as “measurement theory”) take measurement to consist of developing mathematical structures (most often numerical scales) that map on in some systematic way to qualitative empirical relations (Suppes & Zinnes, 1962). Philosophers in this tradition are primarily concerned with the correspondence between the mathematical relations in mathematical structures used in measurement and the empirical relations they correspond to. To this end, these philosophers have used formal proofs to determine the adequacy and limits of using mathematical structures to describe the empirical world.

Philosophers differ, however, in how they describe the relata, or objects, whose relations mathematical systems are meant to correspond with. The relata can be understood as concrete individual objects, qualitative observations of concrete individual objects, abstract representations of individual objects, or as universal properties of objects (Tal, 2017). Further, theories include different adequacy requirements for establishing a mapping between a qualitative empirical structure made up of relata (and the relations among relata) and a numerical scale.

Despite this disagreement, measurement theory has yielded important insights about measurement. First and foremost, the aspects of a mathematical structure used as a measurement scale that “mirror relevant relations among the objects being measured” are the aspects of that structure that are empirically significant. The above example of a mathematical relation, “bigger than” in a length scale is empirically significant because it mirrors, or maps onto, the relation “longer than” between objects. This is and empirically significant aspect in that
knowing the mathematical relation tells us something about the measured system. Therefore, knowing that some object, \( x \), is 8 inches long, while some other object, \( y \), is 4 inches long, and that 8 is larger than 4 tells me that \( x \) is longer than \( y \). Empirical significance can then be used as an adequacy criteria for a measure. To this end, Helmholtz (1887), Holder (1901), and Campbell (1920) argued that expressing a magnitude with a number is adequate “insofar as the algebraic operations among numbers mirror empirical relations among magnitudes.” (Tal, 2017).

### 0.1.2 Measurement theory and psychological measurement

Two claims from measurement theory are particularly relevant to psychological measurement. First is the claim that there is a distinction between representational measures and pragmatic measures. Second is the claim that measurement scales can be categorized based on the empirically significant mathematical transformations they permit.

**Representational v Pragmatic Measures**

Suppes and Zinnes (1962) distinguish between *representational* and *pragmatic* measures. *Representational* measures are those that assign numbers to objects and truthfully represent these objects and their relations. *Pragmatic* measures are those that do not truthfully represent objects, but rather are created with a particular purpose in mind. Natural sciences typically use representational measures, while social sciences more often use pragmatic measures. Length is an example of a representational measure, whereas an economic measure like the Consumer Price Index (CPI) is an example of a pragmatic measure.

Pragmatic measures are common in psychology, where often the thing being measured is a subjective phenomenon like happiness, or well-being. A measure of well-being is pragmatic
both in the sense that the measure is constructed with specific normative implications related to the concept of well-being in mind, and that the measure is constructed with particular uses in mind.

Philosophers following Suppes and Zinnes have argued that the distinction between representational and pragmatic measures is not a dichotomy; many measures have both representational and pragmatic components (Mets, 2019). For example, the Apgar score is a measure of the general health of a newborn baby. The measure is a sum of five factors: 1. skin color, 2. heart rate, 3. reflexes, 4. muscle tone, and 5. breath rate. Each factor is scored on a scale of 1-2. These individual scales for each factors are representational in the sense that they represent features of a phenomenon numerically and the numerical differences (1 vs 2) correspond systematically with physical differences in the infant. However, the Apgar score is pragmatic in the sense that it was constructed, and is used with, a specific purpose in mind: quickly assessing the health of an infant in order to determine whether medical observation or intervention is needed. In order to achieve this aim, the Apgar score sacrifices precision. Rather than use a scale with more values that would more precisely represent biological features, the Apgar score uses a two point scale so that a value can be assigned quickly. Thus the pragmatic necessity of assessing infants as quickly as possible in order to give life-saving medical intervention overrides other criteria.

Throughout this dissertation, I will use intelligence testing as an example of psychological measurement. I argue that history of intelligence testing demonstrate the philosophical issues with psychological measurement. Intelligence tests, like the Apgar score, are both pragmatic and representational. Intelligence scales represent a general cognitive capacity with a numerical score. Differences in scores correspond to individual differences in intelligence. Intelligence tests were created and continue to be revised with pragmatic applications in mind. Different intelligence tests were developed for different purposes, even though all intelligence
tests intend to measure the same ‘thing’. Moreover, the pragmatic applications of intelligence tests has changed over time due to advancements in psychology, shifting societal norms, and political changes.

**Classification of Scales**

The psychologist Stanley Stevens (S. Stevens, 1959; S. S. Stevens, 1951) proposed a classification of measurement scales based on the empirically significant mathematical transformations a scale permits. Stevens’ initial classification distinguished between four types of scales: 1. nominal, 2. ordinal, 3. interval, and 4. ratio. Mathematical transformations are mathematical operations like addition, subtraction, multiplication, etc., performed on numbers. These transformations are empirically significant when they correspond with changes in the measured magnitude, and thus provide knowledge about the magnitude. For example, in a length scale addition is empirically significant because it directly corresponds with changes in the physical quantity. If I add 2 inches to 4 inches, I get 6 inches. In parallel, if I tape a two inch strip of paper to a four inch piece of paper, I now have a piece of paper that is 6 inches long.

Stevens not only created a classification of measurement scales still used today in science, he also was a key player in the debate about the measurability of sensation. At issue was whether numerical scales could accurately be ascribed to sensation intensity. In *Elements of Psychophysics*, Fechner (1860) proposed a method of measuring sensation intensity by recording the ‘just noticeable differences” between sensations of different pairs of stimuli. For example, two sounds of different volumes. ‘Just noticeable differences’ are minimum increments between two stimuli that are noticeable by a subject. Thus the just noticeable difference between two sounds at different volumes is the minimum difference in volumes that is perceivable by a subject. When these differences are assumed to be equal increments of
intensity of sensation, a stable linear relationship exists between the intensity of sensation and the logarithm of the intensity of the stimulus. In other words, if the variation of a stimulus is represented as a geometric progression (i.e. multiplied by a fixed factor), then the variation of a subject’s perception of the stimulus can be represented as an arithmetic progression (i.e. added by a fixed factor). Feschners argued that this relationship, now referred to as “Fechner’s law” (Luce & Suppes, 2002, p. 11-12), provided a method for indirectly measuring sensation intensity via measuring the intensity of the stimulus.

Those measurement theorists, like Campbell (1920), who took the operation of empirical concatenation as a necessary condition for measurement, argued that sensation could not be measured. Concatenation is an operation by which numbers (or other mathematic objects) are joined, end-to-end. Taping two pieces of paper end-to-end is an example physical concatenation. Campbell (1920) argues that concatenations shares structural features with the mathematical operation of addition. These theorists proved that ordering and concatenation are sufficient empirical criteria for additive numerical representation of the target magnitude. While Feschners’s law demonstrates how sensations can be ordered, it does not show how sensations can be concatenated. Because sensation intensities cannot be concatenated (a sensation cannot be added to another sensation in the manner that lengths can), Campbell argued that sensation could not be measured.

Stevens, on the other hand, did not believe measurement requires empirical concatenation. He defined measurement more broadly than Campbell, as the “assignment of numerals to objects of event according to rules” (1951, p. 1). Thus, as long as assignment of numbers to sensation intensity was consistent and not random, it was measurement. For Stevens, measurement did not require that all relations about the numbers in the scale mirrored empirical structures. Some measurement scales could be arbitrary formal schemas justified by their usefulness in describing empirical data rather than by having the proper matching relationship to the
measured system. Sensations can be measured then, even though “sensations cannot be separated into component parts, or laid end to end like measuring sticks” (1975 p. 38).

0.1.3 Psychological Measurement

As the debate about the measurement of sensation demonstrates, measurement theorists focused on the issue of whether measurement of psychological quantities was possible. This dissertation starts with the assumption that psychological measurement is possible and asks, how are measurement methods developed?, and against what criteria are they judged? Even when one assumes the possibility of psychological measurement, the creation of psychological measures is still philosophically interesting.

Psychological measurement is hard. This is in part due to the potentially circular reasoning required to justify a measurement method. The difficulties in measurement justification are exemplified by early intelligence testing. Intelligence is a psychological attribute that is not directly observable. What are observable are a variety of behaviors and success indicators that psychologists hypothesized were related to one’s intelligence (Boring, 1950). These observable criteria, like sensorimotor discrimination, physical strength, or academic scores, served as early, indirect measures of intelligence. However, these measures required that psychologists assumed a relationship between these observable criteria and intelligence. Testing the existence, and nature, of such a relationship was only possible if the researcher already had a measure of intelligence. Thus, the question of “How do we measure intelligence?”, and “What is the nature of intelligence?” are questions that circularly presuppose on another.

Tal (2013) refers to this circular reasoning as the coordination problem.
The problem is that the empirical adequacy of the theory or model and reliability of measuring procedures appear to presuppose each other in a circular way. (p. 2)

The coordination problem appears in the validation of any scientific measurement; however, I will argue that the coordination problem in psychology is especially hard to escape. In the first chapter of this dissertation, I argue that three features of psychology make the coordination problem particularly pernicious: 1. The reliance on single measures, 2. The instability of constructs, and 3. The incomplete (or entirely absent) knowledge of the relevant causal mechanisms. The coordination problem in measurement is a specific instance of circular reasoning common to scientific research. Other forms of reasoning has been examined by philosophers of science, like data-technique circles (Culp, 1995), and the extrapolator’s regress (Steel, 2008). These philosophers have offered up solutions to the circular reasoning, but these solutions are insufficient for the coordination problem in psychology due to these three features.

Psychologist’s use the term “construct validation” to refer to practice of justifying an interpretation of a measure. In the second chapter, I examine the history of “construct validity” in psychology, looking at the various accounts of construct validation. These accounts are useful insofar as they are potential solutions to the coordination problem. However, the majority of the definitions psychologists offer are inadequate as they conflate two related notions: 1. the adequacy of measures, and measures of a construct, and 2. the adequacy of a construct, as a theoretical entity. I argue for a definition of construct validity that respects this distinction. Further, I demonstrate the necessity of this distinction with two examples from research psychology in which measures have a limited degree of adequacy despite the fact that they measure relatively illegitimate constructs.
There are three general approaches to measurement validation that I evaluate in the remainder of the dissertation: 1. operationalism, 2. hypothesis-testing, and 3. coherentism. Both operationalism and hypothesis-testing fail because they do not accurately account for the bidirectional relationship between measure and construct.

In chapter three, I evaluate operationalism, a view initially formulated by Bridgman, Bridgman, Bridgman, Bridgman, and Physicien (1927). Strict operationalists, inspired by logical positivism, claim that a construct’s meaning is nothing more than the operations by which it is measured. For operationalists, the question “What is $x$?” is answered by a operational analysis that answers the question “How do we measure $x$?”. Bridgman’s view suffers from the same problems as the logical positivist philosophy he draws from. While there are many versions of operationalism that attempt to avoid the problems with strict operationalism, the view ultimately fails because it in principle does not distinguish between definitions of constructs and their measures. Thus, operationalism cannot account for revisions of measures or theories once the initial measurement method is determined.

In chapter four, I evaluate hypothesis-testing, a view proposed by Cronbach and Meehl (1955). Drawing from Hempel’s (1952) account of a nomological network, Cronbach and Meehl (1955) argue that construct validation is essentially a process of using a measure to test predictions derived by the theory of the measured construct. Hypothesis-testing, unlike operationalism, permits theoretical constructs that are not exhaustively defined by their measures. However, hypothesis-testing cannot accommodate construct revision driven by technological advances in measurement that develop independent of the theory of the construct. This bottom-up measure revision is exemplified by Spearman’s (1904) factor analysis of intelligence measures. I argue that Spearman’s (1904) revision lead to theory revision, but was itself not dependent on any general intelligence theory.
Rejecting operationalism and hypothesis-testing, I turn to a more recent contender from the philosophy of measurement in chapter five: *coherentism*. Coherentism, as formulated by Chang (2004) and more recently by Van Fraassen (2008), is the view that measures and constructs are validated in tandem, through a process of mutual refinement. The solution to the coordination problem is to shift the focus from the validity of a measure at a single point in time to validation over time. Further, coherentism accommodates the bidirectional relationship between construct and measure. It is not just the case that changes in theory cause changes in measurement, but also that changes in measurement can cause changes in the theory.

The primary limitation of coherentism is that it relies on a vague notion of scientific progress. Revisions of measurement methods or constructs are justified if they progress the theory by increasing some epistemic virtue. Through my presentation of two research projects: implicit memory and intelligence testing, I demonstrated that even a specific epistemic virtue like unification operates differently across psychology. In implicit memory research, psychologists have prioritized convergence between neuroimaging data and their psychological theories of implicit memory. This epistemic goal guides them to revise their functional analysis of implicit memory to reflect the constraints placed by neuroscience. In intelligence research, on the other hand, psychologists have prioritized convergence between intelligence test scores and predictive criteria, so that intelligence tests remain useful diagnostic tools across a variety of contexts.

This points to a need for philosophical work, not about the epistemic virtues themselves, but how they apply differentially across fields. In addition, for epistemic iteration to have prescriptive strength, philosophers need to go beyond mere description of the notions of epistemic virtues, and by connection, epistemic goals, at play. I conclude the dissertation

11
with a brief sketch of how one such account of coherence, Thagard’s *explanatory coherence theory* (2007), could offer constraints for potential epistemic goals.

### 0.1.4 Terminology

Throughout this dissertation I use terms that have a particular meaning when employed by psychologists that is distinct from their meaning in philosophical contexts. Some clarification of terms is necessary, in particular my use of the term “construct”.

Construct-terms, like “implicit memory”, “narcissism”, or “intelligence”, are used in psychological discourse, the nature of what the construct *is*, and thus what the term *refers to*, varies. Slaney and Garcia (2015) identify three categories that uses of the term “construct” fall into: 1. Conceptual, 2. Objective, and 3. Focal phenomenal (p. 246). When used *conceptually*, a construct is a constructed abstraction that includes models, theories, and/or hypotheses about the construct. When used *objectively*, a construct is a mental entity, what Slaney and Garcia call “real-but-unobservable” (p. 246). Finally, when used focally, a construct is an observable phenomenon. In psychology, the focal phenomena are often behavioral patterns.

Thus, Psychologists use the term “construct” to refer to a category of theoretical terms, the entities (mental, cognitive, neurological) these terms are intended to reference and/or track, and a theoretical definition (or description of a model) that contains the term (Slaney & Garcia, 2015). For example, the term “implicit memory” can refer to the cognitive capacity, a theoretical definition of implicit memory that posits various distinctions and causal relations between this and other memory capacities, or to the literal theoretical term “implicit memory”.

1 Adding further confusion, when a term like “implicit memory” is used with the intent of referring to a specific entity, that object can be either a cognitive capacity or a behavioral

---

1 Slaney and Garcia (2015) call the tendency of psychological discourse to use a construct term both conceptually and objectively in the same context the *concept-entity conflation*. 

12
phenomenon. When “implicit memory” refers to the former, its use entails an implicit theory of memory that is not implicit when it refers to the latter.

Slaney and Garcia’s categories are useful in distinguishing between constructs used as concepts, and constructs used as entities. However a further use-mention distinction is necessary (Quine 1940). When a word is used, it is used for its meaning. When a word is mentioned, it is used as a lexical unit. For example, the statement, “Implicit memory is a memory capacity that is distinction from explicit memory”, uses the construct-term “implicit memory”, whereas the statement, “The term ‘implicit memory’ was first introduced by memory researchers in the 1980’s” mentions the construct-term.

For the purposes of this dissertation, a construct is a theoretical term, like “implicit memory”, “intelligence”, or “narcissism”. As such, constructs refer to entities, like cognitive capacities or mental attributes, and are part of psychological theories. Thus, when I refer to the construct of implicit memory, I am mentioning a term independent of a particular memory theory. Different memory researchers may disagree about the best model of implicit memory. It is hard to write about theoretical disagreement in psychology without keeping constructs and theories distinct. I will use either the phrase “construct of x” or quotations around the construct-term, e.g. “implicit memory” when I intend to mention the construct-term, independent of theory.

When I use construct-terms, I use them conceptually. Thus, the statement “implicit memory is an illegitimate construct” is a claim about implicit memory theory. For the sake of clarity, constructs will never be used objectively. Instead, I use the term “entity” to refer to the referents of constructs. Most of these referents will be mental entities, like cognitive capacities or attributes. Constructs can also refer to mental disorders, like depression, or statistical variables, like g.
As the previous discussion has indicated, philosophical definitions of “measurement” vary. For my present purposes, a measure is defined as a set of empirical procedures that are used to interact with a concrete system in order to represent some aspect of that system in abstract terms (e.g. a numerical scale that represents a quantity concept with numbers).

Discussion in psychology about construct validity, and construct validation, often uses phrases like “the measure of a construct” or “construct’s measure”. Technically speaking, it is incorrect to say that a construct itself is measured, as a construct is a theoretical term. Rather, a measurement method measures some aspect of the entity a construct refers to. For the sake of brevity and in order to align my work with psychology, talk of measuring a construct should be understood as shorthand for measuring an aspect, attribute, or property of the entity itself. Thus an “implicit memory measure” should be understood as a measure of the cognitive capacity “implicit memory” refers to.

One final terminological note. ‘Validity” has a technical meaning in philosophy, specific to formal logic. However, in psychology the term has a different, more colloquial meaning. For psychologists, “validity” refers to the property of being justified. Thus, “external validity” refers to a property of generalizations from an experiment, such that an externally valid experiment is one for which generalizations of the findings to non-experimental contexts is sufficiently justified. I will always use the term “valid” in the psychological sense, and not the formal logic sense.
Chapter 1

The Coordination Problem

1.1 Introduction

The success of psychology as a scientific field depends on researchers’ ability to create and validate measures of mental attributes and/or capacities that are often not directly observable. A paradigmatic example of such an attribute is intelligence. Intelligence tests are almost as old as research psychology itself, dating back to Sir Francis Galton’s 1883 publication of a series of articles titled “Inquiries into Human Faculty and Its Development”. The history of intelligence testing is indicative of measurement in psychology more generally. The long history of intelligence testing has produced a plethora of measurement methods and data sets with sample sizes larger than most other subfields in psychology. Most modern intelligence tests in use today have been revised multiple times. The Wechsler Adult Intelligence Scale (WAIS), first published in 1955, is in its fifth edition (Wechsler, 2008). The fifth edition of the test is already underway and will be published next year. Despite multiple revisions, and the large data sets against which these tests are calibrated, modern intelligence tests
have still not managed to escape criticisms that were levied against the first wide-spread intelligence tests.

The most common criticism is that intelligence tests are culturally biased against minorities. A 2005 study by Verney, Granholm, Marshall, Malcarne, and Saccuzzo found that the WAIS-R (a version of the WAIS) was culturally biased such that it had weaker predictive validity for Mexican American students. Concerns that bias in the intelligence tests used in public schools for class placement were causing African American children to be underrepresented in gifted programs and overrepresented in special ed. programs led the California Supreme Court to ban the use of intelligence test for diagnostic purposes on any African American students (Larry P. v Riles, 1979).

Chang’s (2004) treatment of the history of temperature paints an markedly more optimistic picture of measurement validation in the physical sciences. During the early stages of temperature measurement, there was significant disagreement between scientists about the right methods of measurement, including debates about the appropriate scale and about the particular liquid to use in thermometers. Over time, these debates were resolved as scientists were able to determine what environmental factors influenced thermometer readings and settled on mercury as the most reliable liquid.

One may think that due to the age of intelligence testing, we would similarly see a gradual decrease in the disagreement between researchers. Certain empirical phenomena, like the positive manifold demonstrated by cognitive tasks, are well established (A. Jensen, 1998). However, intelligence researchers offer up different, conflicting, explanations of these phenomena. Spearman (1904, 1927) proposed a general intelligence ability, $g$, to explain the positive correlations between tests. Many intelligence researchers still utilize some general intellectual ability to explain individual differences (e.g., Deary, 2002; Deary, Strand, Smith,
Those who interpret $g$ as a psychometric construct only understand $g$ as a statistical factor that summarizes data. On this interpretation, $g$ is a literal mathematical construct, that does not track or refer to a mental entity (i.e. cognitive ability) with a causal role. Even among those who understand $g$ as referring to a cognitive ability, not merely a statistical factor, there is further disagreement about whether $g$ refers to a higher-order cognitive capacity that cannot be reduced to lower-level capacities (like attention, working memory, sensory discrimination), or biological factors (like brain size, neural efficiency, or neural plasticity) (Gray & Thompson, 2004). Thus despite individual differences in intellectual ability being one of the oldest phenomenon of interest in psychology, there is no sign that these debates are closer to being resolved.

Further, intelligence testing is not unique; psychology is littered with examples of measures that while used widely are subject to ongoing criticisms that put the very legitimacy of the measures in question. A more recent example is the Implicit Association Test (IAT), a test that purports to measure a subject’s implicit attitudes by comparing reaction times in a sorting task between two categories of stimuli. The IAT quickly became one of the most utilized measures by social psychologists after its introduction in 1998 by Greenwald, McGhee, and Schwartz. They offered up the IAT as a relatively easy, cheap, and fast measure of implicit bias that was impervious to cheating or lying unlike the self-report and behavioral measures previously used in stereotype research. The IAT became so widely used that it even made its way into the popular media, where journalists touted the measure as proving that “we’re all racists at heart” (Wax & Tetlock, 2005).

Despite its pervasive use, the IAT has been heavily criticized. Some critics believe that the IAT has little predictive or explanatory power; the IAT is only able to predict at most 5.5 of
the variance in an individual’s behavior (Greenwald, Poehlman, Uhlmann, & Banaji, 2009), however, some estimates are less than 1% (Forscher, Mitamura, Dix, Cox, & Devine, 2017). Though the test’s creators have rebutted that the IAT has predictive power for society-wide effects even if the effect sizes are small for individuals’ behavior (Greenwald, Banaji, & Nosek, 2015; Greenwald et al., 2015, see also Payne, 2017). They argue that small effect sizes in the lab can have substantial and wide-spread effects in society at large. Critiques have further charged that the IAT fails to meet basic reliability standards. They claim that IAT is surprisingly unreliable as a measure of an individual’s implicit attitudes (an averaging across all available test-retest reliability measures in 2009 report a median $r = .56$, with a range between $r = .5$ and $r = .6$) (Nosek & Smyth, 2007).

The casual observer may wonder, why is measurement in psychology so hard? This chapter aims to explain why this essential scientific task, measuring the phenomenon of interest, is such a difficult task for psychologists. The answer to this question is evident in one of the first empirical investigations into intelligence. Galton’s early attempts at measuring individual differences in intelligence demonstrates a pernicious philosophical problem in the measurement validation. Efforts to validate a measure of an attribute like intelligence require presupposing parts of the theory of the attribute that the measure itself is intended to test. The reasoning process is circular, such that researchers must first assume the truth of a hypothesis in order to test the hypothesis empirically. Tal (2013) calls this circular reasoning the coordination problem:

The problem is that the empirical adequacy of the theory or model and reliability of measuring procedures appear to presuppose each other in a circular way. (p. 2)
The coordination problem appears in the validation of any scientific measurement. I will argue that the coordination problem in psychology is especially hard to escape. In the first section of this chapter, I will examine the early attempts by Galton to measure intelligence via sensorimotor, physical, and behavioral measures. Galton’s attempt to validate his mental test demonstrates how this circular reasoning is unavoidable, regardless of how successful one’s experiments are. I will then situate the coordination problem within the broader philosophical literature on data-technique circles in section two. Doing so not only demonstrates how the coordination problem is a subtype of this larger epistemic problem in experiments, but also why the problem in psychology is particularly damning to measurement validation. I will focus on three features in particular that contribute to the coordination problem in psychology: 1. The reliance on single measures, 2. The instability of constructs, and 3. The incomplete (or entirely absent) knowledge of the relevant causal mechanisms. These features taken together create a perfect storm in which psychology is currently unable to employ many of the methods used in other research fields to escape data-technique circles.

1.2 Galton’s Intelligence Battery

Experimental psychology emerged as a scientific discipline in the 19th century only when researchers developed methods of measuring the mind with the precision and accuracy that had been achieved in the physical sciences (Fuchs, 2003).

Early research psychologists realized they needed regimented, ‘objective’ measurement methods that could be replicated across laboratory settings. In James Cattell’s influential paper, “Mental Tests and Measurements”, he writes:

Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of experiment and measurement. A step in this
direction could be made by applying a series of mental tests and measurements to a large number of individuals. The results would be of considerable scientific value in discovering the constancy of mental processes, their interdependence, and their variation under different circumstances. Individuals, besides, would find their tests interesting, and, perhaps, useful in regard to training, mode of life or indication of disease. The scientific and practical value of such tests would be much increased should a uniform system be adopted, so that determinations made at different times and places could be compared and combined. (J. Cattell, 15)

Wilhelm Wundt (1832-1920) is widely credited as the father of modern psychology due to his success in developing measurement procedures that were objective (in the sense that they were not based on personal interpretations of the researcher) and capable of being replicated (Fuchs, 2003). Wundt was trained in physiology, which informed his efforts to develop psychology as a natural science (Bringmann, Balance, & Evans, 1975). Using a calibrated pendulum he called a “thought meter”, Wundt created a measurement procedure that enabled him to measure the difference between a subject’s perception of the pendulum’s position and its actual position (Blumenthal, 1985, p. 29). He believed that measuring the difference between observed and actual positions was a means of quantifying a subject’s swiftness of thought. In 1879, Wundt created the first psychological laboratory in Germany, where he administered his thought meter with the intention of explaining individual differences in astronomical observations (Gregory, 2004, p. 5).

It did not take long for psychologists to turn their attention to measuring general mental capacity, i.e., intelligence. Edwin Boring (1886-1968), an influential psychologist in his own right, cites Sir Francis Galton (1822-1911) as the creator of the mental test movement (1950). His publication of Inquiries into Human Faculty and Its Development (1883) argued for two crucial claims. First, that there existed individual differences in mental faculties, and second that these differences could be objectively measured. Taken together, Galton’s claims established intelligence as a phenomenon worthy of scientific inquiry.
Galton developed his ‘objective’ measures of intelligence by adapting the already existing, but still time intensive, procedures used by Wundt into a set of much quicker measures of sensorimotor faculties (Gregory, 2004, p. 6). By streamlining the measurement procedures even further, Galton was able to collect data on more subjects and measure more additional attributes. During the International Health Exhibition that took place in London in 1884, Galton set up a laboratory where he collected measurements on upwards of 17,000 individuals (Johnson et al., 1985). Besides the sensorimotor discrimination tasks, Galton also measured physical and behavioral attributes. A non-exhaustive list of attributes he measured includes: height, weight, head length, head breadth, arm span, length of middle finger, length of lower arm, strength of hand squeeze, vital capacity of lungs, visual acuity, highest audible tone, speed of blow, and reaction time to visual and auditory stimuli. Collectively, sensorimotor discrimination tasks (for brevity these tasks will be referred to as reaction time, or RT tasks) made up what Galton intended to be a comprehensive “mental test” that would allow him to measure individual differences in intelligence and their potential relation to his physical measurements.

An important historical note needs to be added about Galton’s work on heredity and eugenics. Charles Darwin was Galton’s cousin, and Darwin’s publication of The Origin of Species in 1859 shaped the rest of Galton’s research. Galton was primarily interested in what human ability was hereditary (Fuchs, 2003, p. 12).

He advanced the view that mental capacities were inherited in a book called Hereditary Genius, in 1869. This work laid the foundation for eugenics. Eugenics is now considered an illegitimate and immoral view of human heredity based on a misunderstanding of evolution. However, eugenics is relevant to understanding Galton’s motivations when he set up his lab at the International Health Exhibition. Galton was interested in individual differences in both sensorimotor discrimination and physical attributes because that would serve as evidence for
his views about heredity. His goal was to discover any traits or abilities that may vary across the population, as these would be relevant variables for his studies of heredity. A direct consequence of these views is that Galton held, as a background assumption, that mental and physical traits were distributed normally across the population and were positively correlated with each other (D. J. Galton & Galton, 1998).

For our present purposes, I will focus on Galton’s use of RT as an indirect measure of mental capacity (what would now be referred to as intelligence). This early example of intelligence testing demonstrates a pernicious philosophical problem underlying psychological measurement. Galton’s interest was not in individual differences in RT themselves. Instead, he was interested in RT in so far as sensory discrimination measures could support inferences about the mental trait he could not directly measure: intelligence. Galton’s measurement method involved using direct measures of things he could observe in order to ‘get at’ something he could not observe.

Using measures of observable quantities as an indirect measure of non-observables quantities is the foundation of modern psychological measurement. To this day, psychologists depend on indirect measurement methods where they infer a non-observable trait or capacity from measures of observable phenomenon. Reaction time to stimuli, first used by Galton and Wundt, is still an important measurement in psychology. Now commonly referred to as ‘response time,’ or RT, measures of RT are used to measure everything from a subject’s attention to their implicit bias. For example, the IAT discussed above directly measures a subject’s response time in a sorting task, and infers their implicit bias from a comparison of the subject’s average RT between two categories of stimuli (Greenwald et al., 1998).

Despite the pervasiveness of this kind of indirect measure, Galton’s mental test demonstrates how the logic of indirect measurement is potentially circular. Galton’s goal was to infer a
subject’s intelligence, an attribute he could not directly measure, from the data produced by measuring things he could directly measure, sensorimotor discrimination. However, Galton did not know the exact relation between the attributes he could measure and the attribute he wanted to infer. While from our modern viewpoint, using RT to measure intelligence, albeit indirectly, is counterintuitive, Galton was not alone in thinking that mental capacities were related to sensorimotor discrimination. Wundt’s thought meter rested on a similar assumption: that speed of thought (a mental attribute) could be inferred from a subject’s accuracy at a visual sensorimotor task. Both Galton and Wundt believed that sensorimotor discrimination, and also physical attributes like strength and height, were positively related to intellectual capacity.

While Galton was not alone in believing a relationship existed between his measured attributes and intelligence, the only method available to him to justify this belief was to test it empirically. Here in lies the potentially vicious circularity. Any empirical test of the hypothesized relationship between intelligence and RT would require a measurement method of both variables. If Galton could directly test the relationship between the observable and non-observable variables, he could provide evidence that could independently validate his measurement method as a legitimate measure of intelligence. However, the only measurement method available to Galton was the test battery he had developed. Thus the only way Galton could test the assumption on which his measurement method was founded would be to use the very measurement method in question. From his position, there was no independent measurement method to test the hypothesized relationship. It seems that Galton could only validate his measurement method by begging the question and assuming his measurement method was a legitimate measure of intelligence.

Tal (2013) calls this type of circular reasoning in measurement the coordination problem:
The problem is that the empirical adequacy of the theory or model and reliability of measuring procedures appear to presuppose each other in a circular way. (p. 2)

Galton’s coordination problem is that he has two questions, “What is intelligence?”, and, “Does this RT measurement procedure measure intelligence?”. Answering either question requires assuming the existence of the relationship between the observed and inferred attributes. To answer this first question Galton would also have to answer the question of whether a relationship existed between sensorimotor/physical attributes and intelligence. However, to answer the second question, Galton would also have to answer the relationship question. He can only determine the validity of his measure by testing it. In employing his mental test, Galton has to assume the relationship exists. These two questions are inextricably linked by this shared implicit question about the relationship between the observed and inferred attributes. Galton cannot answer either question without presupposing an answer to the other. But in doing so, he has assumed the answer to the question he was making the assumption in order to answer. Alexandrova and Haybron (2016) describe the coordination problem as “something of a chicken-and-egg-problem in that, if we already knew exactly what the correlations a measure should exhibit, we might not have much need of the measure.” (p. 5).

1.2.1 Formalizing the problem

Galton’s coordination problem generalizes into a more general epistemic problem that applies to any case in which one attempts to infer an unobservable quantity from an observable quantity. Galton wants to know x (a subject’s intelligence), but only knows y (a subject’s reaction time). He attempts to infer x from y, all the while not knowing what the relationship between y and x is, or if one exists at all. Thus Galton’s experiments rest on the assumption
that $x$ is a function of $y$, $x = f(y)$, i.e., intelligence is a function of sensorimotor capacities and physical attributes. The nature of this function is unknown. Even more troubling, the only way Galton could determine if such a relationship exists, and the nature of the relationship, is to use his RT measure. Of course, using the measure itself requires assuming that the RT-intelligence relationship exists, thus forcing him to presuppose the relationship that he intends to test. It seems that measuring a quantity that is not directly observable (in this case a mental capacity, intelligence) leads to a vicious circularity, in which one has to assume a relationship exists between two things in order to determine if a relationship exists between two things. Hassock Chang (2004) refers to this problem as “the problem of nomic measurement”, and formalizes the problem as follows:

1. We want to measure quantity $X$.

2. Quantity $X$ is not directly observable, so we infer it from another quantity $Y$, which is directly observable.

3. For this inference we need a law that expresses $X$ as a function of $Y$, as follows: $X = f(Y)$.

4. The form of this function $f$ cannot be discovered or tested empirically, because that would involve knowing the values of both $Y$ and $X$, and $X$ is the unknown variable that we are trying to measure. (paraphrased from Chang, 2004, p. 59).

Chang’s formulation here highlights the fact that coordinating between the measured and inferred quantities requires positing a law that connects the two in a mathematically describable way. However, the problem is not limited to scientific theories that posit (or aim at positing) laws, or even law-like generalizations. We can substitute “law” in (3) with any type of scientific theory representation, like a model or causal mechanism, that represents
one variable as a function of another. The problem arises in any situation in which measurement depends on positing a relationship between observable and non-observable quantities, regardless of how that relationship is represented in the relevant scientific theory.

Further, the coordination problem does not require a clear distinction between observable and non-observable quantities. Chang’s formulation invokes a distinction between directly observable quantities, which are properties we can measure, and non-directly observable quantities, which are properties that must be inferred from what we can directly measure. This could be an issue, as observability is notoriously hard to define. If we cannot systematically distinguish between which quantities are observable, and which are not, one may argue that we are incapable of diagnosing coordination problems. Alternatively, if one thinks that nothing is directly observed in science, then it would seem that the coordination problem is ill-formed. Rather than inferring non-observable quantities from observable quantities, researchers infer non-observable quantities from other inferred quantities.

Carnap provided perhaps the clearest account of observability (1966, ch. 23). On his account, a property is observable iff 1. Perception of the property is unaided by instruments, and 2. Perception of the property is unaided by inferences. However, even Carnap did not believe the distinction was clear; instead he thought that observability was a continuum of degrees and potentially subject-dependent.

Carnap’s definition of observability for philosophers is not, however, what Chang has in mind. On Carnap’s account, something that is directly measurable is, de facto, not observable, as the perception of the property requires the use of an instrument. What Carnap calls the “scientist’s sense” of observability is the relevant sense of observability. The scientist’s notion of observability is broader than the Carnapian one and permits the use of instruments to aid observation.
However, categorizing instrument-aided observation as direct observation raises an additional set of worries about observability. Understanding observability this way means that what is observable can change over time, as new instruments are invented and become more sophisticated. For example, it seems that while electrons were not observable prior to the invention of the cloud chamber, they are/were observable after (Achinstein, 1965).

The best response to the thorny issues raised by the observability distinction is to side-step the distinction altogether. Like with law-like generalizations, while Chang’s formalization of the coordination problem invokes the observability distinction, it does not actually require that a clear distinction between observable and non-observable quantities exists. All that the coordination problem requires is that there exist some quantities that are inferred from the measurement of other quantities. Neither the measured quantity nor the inferred quantity may be directly observable, depending on one’s particular notion of observability. However, both quantities must differ in their degree of observability, such that the measured quantity is more directly observable than the inferred quantity.

Chang’s formalization is helpful because it illuminates that the ‘circularity’ Tal (2013) refers to is caused by making presuppositions about the very theory that the measurement procedure is intended to test. Understood this way, the coordination problem is not the result of an infinite regress. Rather, the coordination problem is one version of a much broader problem in the philosophy of science: data-technique circles. Data-technique circles involve reasoning about data that requires making assumptions about the very theory one is testing. This circular reasoning is problematic to philosophers who want to defend the claim that science is, or aims at being, objective. Efforts by philosophers to break data-technique circles are helpful, in so far as they point to possible non-circular sources of evidence for measurement validation. However, these methods are also troubling for psychology because they require knowledge standards that psychology cannot currently meet.
1.3 Data-technique circles at large

Data-technique circles are a type reasoning about empirical data in which researchers must assume, at least in part, the truth of the theory/model which they intend to test. This type of circular reasoning has received extended treatment from philosophers of science. Here, I will focus on three formulations of data-technique circles that are particularly relevant to the coordination problem in psychological measurement. First, there is Collins’ (1985) discussion of what he calls “the experimenters’ regress”. The experimenter’s regress refers to how experimental practices in their entirety, including the interpretation of data, require circular reasoning that can only be broken with non-scientific reasons. Second, there is Culp’s (1995) argument that the experimenter’s regress, which she refers to as data-technique circles, can be broken by employing a notion of robustness. Finally, Steel’s (2008) explanation of the extrapolation problem in the social sciences reveals an additional worry about psychology in particular. Steel’s work on extrapolation shows how difficult it is to justify generalizations of experimental findings in the social sciences in which the sample populations are chosen often not because they are representative of the general population but because of practical limitations.

I will argue that the coordination problem is one instance of a data-technique circle. As such, the solutions offered to data-technique circles, in particular by Culp and Steel, may be relevant to formulating a solution to the coordination problem. Their solutions point to different sources of evidence that a researcher could use to in effect ‘bootstrap’ her way out of circular reasoning about her measurement method. Culp (1995) argues that researchers must utilize a notion of robustness, in which data from one measurement method is strengthened if it is similar/concordant with data produced by an independent measure. Steel (2008)
describes a process tracing reasoning process by which researchers justify generalizations of empirical findings with knowledge of the underlying causal mechanism.

While these formulations of data-technique circles illuminate various ways in which measurement validation can be circular, the solutions offered are ill-suited for the coordination problem in psychology. This is because Culp’s and Steel’s solutions set criteria that most of psychology cannot satisfy: robust evidence. Robust evidence is evidence that is arrived at using different empirical methods that hold different theories/theoretical claims as background assumptions. If validating psychological measures requires that evidence is robust, current psychological research projects are likely unable to meet this criterion due to their reliance on single, paradigmatic measures. Steel’s solution to the extrapolation problem requires a well-developed understanding of relevant causal mechanisms. In psychology, this might be too tall of an order, considering that most constructs in psychology are loosely defined such that the theory is ambiguous between multiple different potential mechanisms. I conclude that while data-technique circles and the extrapolation problem deepen our understanding of the variety of ways in which measurement validation involves theoretical presuppositions, they do not resolve the coordination problem itself. Instead, they reveal why measurement validation in psychology has proven to be so difficult.

1.3.1 The Experimenters’ Regress

Collins (1985) defines the experimenters’ regress as a problem that arises from the circularity between experimental techniques and the data produced by these techniques (p. 84). Experimental techniques are the collection of measurement and intervention procedures by which researchers empirically test the hypothesis of a particular theory. Collins argues that no experiment is a truly objective test of a hypothesis, because experimental techniques are not truly replicable (p. 19).
Collins’ argument is as follows. A researcher determines whether or not a particular experimental technique is ‘working’ by establishing that the technique produces the expected, i.e. ‘correct,’ data. This creates a circular regress between experimental techniques and the data they produce: a regress in which a researcher trusts the data produced by an experiment because a reliable instrument produced it, but the instrument is judged to be reliable only because it produces the right kind (e.g. ‘correct’) of data. The only way to break the experimenters’ regress is to resort to “‘nonscientific tactics’ . . . because the resources of the experiment alone are insufficient” (p. 143). These nonscientific tactics are appeals to reasons external to the empirical data. Collins, a sociologist of science, is primarily concerned with socially motivated nonscientific tactics, like the social interests of his scientific community, or a scientist’s career ambitions. The social factors that motivate a researcher to interpret their data in a particular way, and enable them to break the experimenters’ regress prevent true replication of experiments. These non-scientific motivations will be unique to the individual researcher, and thus a different researcher who may use the exact same protocol, will differ in their use of nonscientific tactics and interpret the data differently. As replicability is the demarcation criterion for objective knowledge (p. 19), experiments are not objective tests of hypotheses.

The experimenters’ regress as Collins presents it is intended to motivate his larger point about the use of nonscientific tactics in scientific reasoning. Collins followed the work of Feyerabend (1962) and T. Kuhn (1962), who argued that all scientific observation is theory-laden, in that all scientific observation requires making theoretical commitments.

His focus on the lack of objectivity in scientific observation is directly relevant to understanding the coordination problem. Measurement validation is just one particular form of theory testing. As measurement itself requires tacit assumptions about the relevant scientific theory, there is no theory-independent measurement validation. Measurement methods are judged to
be reliable only by assuming the very theory it is intended to test, just as any theory testing requires theoretical assumptions.

In describing the regress, Collins’ aim is to draw attention to how science fails to be objective without necessarily providing a solution. The only route out of the regress is the use of nonscientific tactics. Other philosophers, in particular, Culp (1995), have discussed the experimenters’ regress as a potentially surmountable problem rather than an inevitable source of subjectivity in research.

1.3.2 The Importance of Robustness

Culp (1995) refers to the experimenters’ regress as a data-technique circle in which interpretations of data depend on idiosyncratic presuppositions:

A scientist will fail in making objective interpretations of her raw data to the extent that her interpretations are biased by dependence on idiosyncratic presuppositions. (p. 439-440)

A researcher is biased by these idiosyncratic presuppositions even if the presuppositions are true. For example, Galton could be correct in his assumption that grip strength positively correlates with a subject’s intelligence, thus making the presupposition true. However, Galton could still be wrong about the broader theory that entails this assumption, so that the relationship between grip strength and intelligence is an extraneous correlation that is due to some underlying biological/psychological mechanism that relates physical and mental ‘strength.’

Culp advocates for using a standard of robustness in measurement methods in order to break data-technique circles. Robustness, according to Culp, is a property of data produced by a
measurement method. A set of data is more or less robust depending on the degree to which it agrees with, or matches, data obtained by independent measurement methods (p. 441). Measurement methods are independent if they rely on different theoretical presuppositions.

Independence between measurement methods is often imperfect, meaning that two measures may share some theoretical presuppositions, but differ in others. Thus, robustness itself comes in degrees, depending on both the degree of agreement between the data and the degree of independence between the measurement methods.

The motivating intuition for robustness is that when two independent measures produce the same effect, it seems unlikely that the agreement between the data is purely coincidental. Culp argues that when effects are robust, it is more likely that the observed similarities in the two sets of data are due to some shared underlying cause:

When comparable data can be produced by a number of techniques and the raw data interpretations for these techniques do not draw on the same theoretical presuppositions, this remarkable agreement in the data (interpreted raw data) would seem to be an improbable coincidence unless the raw data interpretations have been constrained by something other than shared theoretical presuppositions. (p. 448)

Culp’s argument relies on the principle of the common cause (Salmon, 1984; H. Reichenbach, 1956). The principle of the common cause as formulate by Rechenbach (1956, p. 157) states that “if an improbable coincidence has occurred, there must exist a common cause.” Salmon (1984) cites Perrin’s use of Avogadro’s number, the number of molecules in a mole of a pure chemical compound, in Les Atomes (1913) as a paradigmatic example of the principle. Perrin emphasizes that Avogadro’s number can be determined in multiple distinct ways. This “remarkable agreement” across all of the independent methods of calculating Avogadro’s number, to Perrin, is strong evidence that there is some common cause (the existence atoms,
ions, and molecules). If molecules did not exist, as a common cause to each independent
determination of Avogadro’s number, “These different experiments designed to ascertain
Avogadro’s number would be genuinely independent experiments, and the striking numerical
agreement in their results would constitute an utterly astonishing coincidence” (ibid., 220).

In other words, the determination of the same value (Avogadro’s number) using independent
methods is an coincidence that is improbable without some common cause underlying each
seemingly independent method. The common cause in this case is the existence of atoms.

While Culp’s focus is on general reasoning about experimental data, her argument is directly
relevant to measurement validation. Judging that a measurement method is reliably tracking
some quantity requires reasoning about, and drawing inferences from, experimental data
produced by the measurement method. Thus the validity of a measurement method can be
thought of as one type of scientific inference that involves data-technique circle reasoning.

Culp’s solution to the coordination problem, then, would be to use the notion of robustness
and compare independent methods of measurement of the same attribute. Galton could utilize
robustness to add external evidence to his interpretation of his experiment by comparing
the data produced by his original test battery to other criteria that relate to his theoretical
definition of intelligence. For example, say that Galton found that not only did his measured
attributes strongly correlate with each other in each subject, but also correlated with criteria
(variables related to a construct that are directly observable), like the subject’s income level, or
academic grades. This would be additional evidence that not only do his measured attributes
relate to one another, but they also relate to intelligence itself.
Stegenga (2009) calls this type of evidence “concordant multi-modal evidence”, and astutely points out that it is difficult to determine if the measurement methods of the original attribute of interest and the criterion are truly independent:

Simply put, having independent modes of evidence and knowing that they are properly independent are difficult, since robustness requires multiple modes of evidence, and incomplete or vague individuation of evidential modes will leave robustness as an incomplete or vague notion, and hence robustness-style arguments will be vague or inconclusive. (p. 653)

Stegenga uses the case of mesosomes to demonstrate his point. Starting in the late 1950s, biologists observed folded invaginations in the plasma membrane of bacterial cells. Mesosomes in bacteria is an example of a phenomenon whose existence was supported by multiple (supposedly independent) methods, that is now understood to be merely an artifact of previous chemical fixation techniques for preparing samples for electron microscopy. These chemical fixation methods damage the cell wall, creating the indent that was interpreted as a structural component of the cell membrane. Even though there were multiple methods for preparing cell samples, indicating at least some degree of robustness, all the previous methods were biased in the same way that lead to a shared “false” effect. This false effect was initially interpreted as evidence for the existence of mesosomes because the effect appeared to be robust. In Culp’s argument for robustness, she claims it is more likely that robust effects are due to a common ‘real’ cause (real in the sense that the cause is not an artifact of the experimental paradigm). However, the mesosoma case demonstrates that it is not impossible for such artifacts to occur.
Analogously, it is an open possibility that Galton’s measurement of sensory discrimination and a subjects’ academic grades are not independent measurement methods. Perhaps one’s sensory discrimination relates to how one performs on a test if that test is timed and involves visual discrimination tasks. Like in the mesosoma case, concordant data would be an artifact caused by the measurement methods used. The data would appear to be concordant and robust only because we are ignorant of the bias in both measures due to their shared reliance on sensory discrimination.

Whether or not measures are truly independent matters in so far as it mediates the strength of data concordance as evidence for a measure’s reliability. Stegenga’s worry about evaluating robustness is two-fold. It is both challenging to develop genuinely independent measurement methods, and difficult to ascertain the independence between measurement methods. Culp’s argument for robustness “is a type of no-miracles argument: it would be miraculous if multiple independent experiments showed x. . . and x was not real” (p. 653). However, if it is difficult to develop independent measurement methods, such that most measures are not entirely independent, and difficult to determine the independence, one has to worry that what would appear to be miracles from the researcher’s perspective become a lot more common.

Stegenga’s criticism of robustness goes even further, to interpretations of discordant multimodal evidence. One may think that if independent measurement methods produce data that disagrees with the original data, a researcher’s ability to determine the robustness of the evidence is irrelevant. Instead, Stegenga argues that we still require robustness for interpreting discordance evidence. If Galton failed to produce evidence for the expected correlations, he still faces an interpretation question that is underdetermined by his available evidence. If the measures do not correlate with each other, there are two possibilities: either the measures themselves are faulty, or the presupposed relationship between the measured attributes does not exist. It could be the case that due to the limits of the available technology at the time,
Galton’s measurement methods were unreliable. Unreliability would introduce random error into his data that could mask the correlations between measured attributes. However, the same effect, or lack of effect, could also be interpreted as evidence that the relationship Galton thought existed between physical attributes and intelligence did not exist. As with concordant evidence, evaluating discordant evidence requires Galton to determine the degree of independence between measurement methods. Failure to do so accurately could lead to erroneously accepting discordant data as evidence that the measure and criterion were not related.

Robustness, while initially promising, does not offer a sufficient solution to the coordination problem. While robustness introduces a new form of evidence, the original issue of employing circular reasoning while interpreting positive and negative evidence rears its head again. Whereas before the problem was that validating the measure required presupposing the existence of a relationship between the observed and inferred quantity, evaluating robustness evidence requires presupposing the independence (or dependence) between measurement methods.

The Independence of Psychological Measures

Criteria are robust evidence for a measure only in so far as your methods of measuring the criteria are independent from the measure being validated. Determining the independence of two measures is problematic not only because robustness is vague, but also because our ability to accurately judge independence is fallible. Measurement methods are independent when they rely on different theoretical assumptions. However, there are two threats to measurement independence in psychology. First, often an entire research field will rely on a single, paradigmatic measure for a phenomenon of interest. Second, when there are multiple measurement methods available, it is often the case that the different measures were in fact
“validated” by agreeing with the already existing measures, and/or that the different measures rely on the same sensory modality or type of stimuli. These measures are hardly independent and thus cannot provide robust evidence for psychological effects.

Research projects that only use a single measurement method fail to meet the Culp’s criteria for robust evidence. However many fields exist in a gray area, in which what appear to be different measurement methods of an attribute or capacity, in fact, share the same logic or use the same modality and thus make the same theoretical assumptions. Further, shared theoretical assumptions are found even across measures that do not measure the same attribute, but instead measure different attributes or related criteria. Intelligence testing is one such case, where different tests often rely on similar tasks. The vast majority of intelligence tests today that are developed for use on neuro-typical adults contain variations of the same tasks created by the authors of the first non-reaction time-based intelligence test: Simon and Bellevue (Boake, 2002).

In 1904, the Minister of Public Instruction in Paris created a commission to determine what should be done to address the educational needs of children with mental retardation. The commission decided that such children needed special instruction, apart from regular schooling, and that a diagnostic test was needed to identify these children. Alfred Binet (1857-1911) and Theodore Simon (1872-1961) produced the first edition of the Binet-Simon Scale in response to the commission’s recommendation in 1905 (Gregory, 2004).

Binet and Simon created a test made up of only higher-cognitive reasoning (e.g., judgment, comprehensions, problem-solving, etc.), rather than tests of sensorimotor discrimination. This shift was part of a more substantial shift away from sensorimotor discrimination tasks after Clark Wissler’s (1901) work failed to find any meaningful correlations between sensorimotor discrimination and academic performance (Boake, 2002).
The test only consisted of thirty items, which were administered individually. In a revision of the test published in 1908, Binet and Simon incorporated a new scoring method consisting of a ratio of mental age to chronological age, the first “Intelligence Quotient” (Gregory, 2004).

Binet and Simon demonstrated the validity of their intelligence scale with evidence that scores on the test increased with age (meaning age and test scores were positively correlated) and that test scores differentiated between normal and cognitively impaired children in agreement with diagnoses performed by independent psychologists (Gregory, 2004). Thus the two most important criteria for the first intelligence test was the previous diagnosis of mental deficiency and chronological age.

Both criteria lead to Stegenga’s worry about interpreting concordant evidence. That chronological age should positively correlate with intelligence is a theoretical assumption about intelligence. Binet and Simon assume a theoretical definition of intelligence by which it was predicted to positively correlated with age. Even though chronological age can be measured by a clearly independent method, that age is a relevant criterion by which to judge the validity of an intelligence test was itself an unjustified theoretical presupposition at the time. It is entirely possible to define intelligence differently; still maintaining much of the folk meaning behind the term, but disconnecting it from chronological age.

The agreement of test scores with previous diagnoses by independent psychologists is also a problematic criterion because the diagnostic methods employed were not theoretically independent from the items in the scale. The diagnostic tests available at the time contained similar higher-order reasoning tasks (Boake, 2002). Edouard Seguin (1812-1880), for example, developed a form board task that requires spatial reasoning to diagnose children (Gregory, 2004). The original Binet-Simon test includes several test items that measure spatial reasoning via tactile tasks like Seguin’s form board, including a paper cutting task. It is entirely possible
then that a child who was previously diagnosed as "feeble minded" and then scored in the "feeble minded" range on the Binet-Simon scale was diagnosed using the same items, or similar items, to those on the Binet-Simon scale. Thus, independent diagnosis is hardly an independent method of measurement and provides at best weakly robust evidence for the validity of the Binet-Simon scale.

The justification for the Binet-Simon scale reveals that what was considered evidence for the validity of early intelligence tests was not robust in the sense that Culp requires. The evidence Binet and Simon appealed to came from measures that made many of the same theoretical assumptions as their test. This is not a unitary phenomenon, either. The intelligence tests that came after the Binet-Simon utilized the same type of evidence and borrowed heavily from already established test items. For example, the WAIS, which is currently the most common intelligence test today, borrowed heavily from the Stanford revision of the Binet-Simon scale and differed from the previous test primarily in how the test was administered and scored (Boake, 2002).

While this is merely one example of failed measurement independence, it is illustrative of the more general problem in psychology. Independence between most psychological measures of the same phenomenon is lacking, to a degree that threatens the plausibility of robustness as a silver bullet for the coordination problem. This is a problem not just in well-established subfields like intelligence research and more recent subfields like implicit attitudes research. While researchers have developed multiple measurement methods of implicit attitudes, they share core assumptions about the relationship between attitudes and reactions times.

To add further doubt to the ability of robustness to solve the coordination problem, the multiplicity of measures often means discordant data. In memory research, for example, different measures of implicit memory rely on different sensory modalities, achieving a
significant degree of theoretical independence. However, implicit memory effects observed are not consistent across these measures. Manipulating a variable using a spatial-reasoning task to measure implicit memory, for example, produces different results than when using a language-priming task (Dew & Cabeza, 2011). Similar problems occur in episodic memory research, where performances on different episodic memory tests do not correlate for both children and adults (Cheke & Clayton, 2015).

Therefore, robustness alone cannot solve the coordination problem in psychology. Many fields of psychology fail to achieve independent measures of the phenomenon of interest. Moreover, when some dependence between measures is achieved, researchers fail to produce the necessary concordant data. Thus, while one may disagree with Stegenga and believe that one can judge the independence of measurement methods without presuppositions, in psychology measure independence that yields robust evidence for the validity of a measure is rare.

1.3.3 The Extrapolation Worry

When we last left Galton and his RT mental test, we determined that validating his test required some additional evidence outside of the data produced by his specific experiment. External evidence has the potential to break the vicious circularity in Galton’s reasoning by justifying his presupposition that a relationship existed between RT and intelligence. However, Stegenga’s worry about evaluating the independence of measurement methods threatens the plausibility of this approach. His worry goes both ways. If the external evidence seems concordant, Galton’s belief that the two measurement methods are truly independent is underdetermined by his available evidence. If the external evidence is discordant, Galton is also unable to determine the independence of the measurement methods definitively.
Stegenga claims that discordant evidence is more common than concordant evidence. He is primarily concerned with the general tendency for researchers to ignore, or even explain *ad hoc*, discordant evidence. The claim is that researchers fail to appreciate discordance when they have enough motivation to (both theoretical and pragmatic) to believe both measures are producing the ‘same’ effect. However, there is another possible type of error in reasoning that discordant evidence can cause. When a potentially independent measure produces discordant evidence, inaccurately judging the independence of the relevant measures can lead to a kind of Type II error. In these cases, a hypothesized relationship between the variables is rejected based on the discordant evidence. This is an error when the two measurement methods are independent in a way that erroneously causes the discordance. While the discordance is real, i.e., the effects differ between the measures due to ‘real’ differences in the phenomena being observed, the observed differences are not in fact due to the relationship between the variables not existing. Instead, the relationship exists but is masked.

Wissler’s (1901) study is an example of premature measurement rejection caused by this kind of error in evaluating discordant evidence. Clark Wissler (1901) was the first psychologist to explore the relationship between RT mental tests directly, like Galton’s, and criteria that were assumed to be related to intelligence using correlational analysis. Wissler collected RT mental test scores and academic grades from a sample of more than 300 students at Columbia University and Barnard College. He predicted, in line with his predecessors, that RT would positively correlate academic performance. However, his results showed virtually no relationship between the variables, and even more astoundingly, an only modest correlation between the subtests of the mental test itself. For example, hand movement speed was positively correlated with color naming (r = .19). When he published his findings in 1901, experimental psychologists largely abandoned RT and more generally any sensory discrimination based measures of intelligence (Gregory, 2004).
This whole-scale rejection of the relationship between RT and intelligence turned out to be a mistake. It was not until more than 70 years later when A. R. Jensen (1982) reported correlations of -.40 between complex forms of RT and intelligence that psychologists realized their error. Counter Wissler’s findings, Jensen provided evidence that there was a relationship between RT and intelligence. The relationship just happened to be the reverse direction.

Looking at potential explanations of why Wissler failed to find meaningful correlations between his mental tests and academic performance points to a related problem with social science research that further deepens the coordination problem. As Gregory (2004) states, “. . . by contemporary standards Wissler’s research methods revealed an extraordinary psychometric na’iveté” (p.8). It is unsurprising that Wissler failed to produce the correlations that A. R. Jensen (1982) later found, considering the lack of representativeness of his sample. Wissler’s sample consisted only of college students at Columbia and Barnard. By sampling only from college students, Wissler created an accidentally homogenous sample of subjects. His subjects were all of the same age, similar socio-economic backgrounds, and had to have achieved relatively high academic grades in order to become students at Columbia. This in effect artificially limited the variance in scores on Wissler’s tests, which in turn restricts any correlation analysis of the data. Wissler’s measurement protocol also involved a small number of trials per subject, which compounded with his sampling issues decreased the reliability of response time scores. Lack of reliability places a severe restriction on the upper bounds of correlation coefficients, as it introduces random error that can mask the effect that is there.

On the one hand, Wissler unintentionally decreased variance in one of the measured variables: academic achievement, and on the other hand unintentionally increased variance in another: response time. Variance is directly related to correlational analysis because correlation is calculated from the covariance of two variables. In other words, the correlation of two variables estimates to what degree these variables vary together. By underestimating variance
in one variable, and overestimating variance in another variable, Wissler effectively hid the relationship between the two variables with measurement error.

All of this amounts to a lack of what psychologists call external validity. External validity refers to the degree to which the results from a study can be generalized to a broader population, i.e., whether the effects in the lab also exist outside the lab. External validity is negatively impacted by artificially homogenous samples, and unreliable measures, like Wissler’s. Wissler’s original sample was not representative of the general population. Thus psychologists were not justified in generalizing Wissler’s findings as they did.

The problem with non-representative samples is not unique to Wissler, early psychologists, or even psychology at large. Steel (2008) discusses the issue of external validity more generally as it applies to any scientific field that studies live populations (e.g., biology, sociology, psychology, etc.). Generalizing from a sample population to a larger population involves a type of inference called extrapolation. When a researcher extrapolates, they start with knowledge of a causal relationship in one population and infer a causal relationship in a different population (p. 3).

Extrapolation is straightforward when the populations in question are perfectly homogenous. This is often the case for the ‘harder’ sciences. A chemist can experiment with a sample of a single atom of hydrogen and be pretty confident that his findings generalize to all hydrogen atoms, assuming the conditions are similar. What makes extrapolation hard for researchers in the life sciences is that even if they want to extrapolate a finding only to conditions that are precisely the same as the conditions in the lab, the population they want to extrapolate to is not homogenous. All hydrogen atoms are (almost) identical. Humans, in comparison, are definitely not.
Steel refers to the difficulty in generalizing to non-homogenous populations as the problem of extrapolation in heterogeneous populations, or, the problem of extrapolation (2008, p. 3). The problem here is that when the population you want to extrapolate is heterogeneous, it is possible that the population differs from the sample population in a way that affects the causal relationship you want to generalize.

Wissler’s study and the premature rejection of a relationship between sensorimotor discrimination and intelligence demonstrate what happens when we incorrectly judge a sample to be representative. Wissler’s sample population differed in just the right way that masked a causal relationship that exists in the general population. While in Wissler’s subjects, there was no relationship between their response time and their intelligence, Jensen’s (1982) study provides evidence that there is such a relationship in the general population. Jensen’s study differs from Wissler’s by having a more representative sample in which the subjects are less homogenous in terms of age and education level.

As is the recurring theme of this chapter, the problem of extrapolation is a problem caused by circular reasoning. The extrapolator’s circle is that extrapolation is only worthwhile when there are restrictions on what a researcher can learn from studying the target population directly. However, these restrictions that limit a researcher’s knowledge about the target population, motivating an extrapolation from a different sample population, limit the researcher’s ability to justify that very extrapolation.

The use of model-organisms in biomedical research is a prime example of the extrapolator’s circle. Researchers who study novel drug interventions for Alzheimer’s use mice as model-organisms because of the ethical and financial concerns with using human subjects. These drugs are dangerous and often expensive, making human testing in the early stages of drug development impractical and highly unethical. Thus there are real restrictions on
Alzheimer’s research on the target population: humans with Alzheimer’s. This necessitates that researchers extrapolate from research from a different population, in this case, mice in which Alzheimer’s like symptoms are induced. However, that same limitations that restrict testing on humans limit our knowledge of the relevant causal mechanisms involved in Alzheimer’s Disease.

For example, researchers currently disagree about whether the overabundance of senile plaques, one of the primary markers of Alzheimer’s on the neural level, is merely a symptom of the disease, i.e., a downstream effect of some underlying neural-biological process, or a primary cause of the disease (LaFerla & Green, 2012). This disagreement matters to extrapolation from mice models to human models because the causal process that artificially causes plaques in the mice is quite different from the process in humans. This difference is a relevant difference between the mice population and human populations only if plaques are a symptom, rather than a cause, of the disease.

Attempts to investigate the role of plaques in Alzheimer’s are limited, however, because of the practical limitations on human research. Which of course is precisely the question that researchers need to answer in order to justify extrapolations from mice models to human models, extrapolation that is necessitated by the same limitation on human research. Once again we are faced with dizzying circular reasoning, where one must assume an answer to a research question in order to answer the question. Thus if we develop a drug that the mice with artificial Alzheimer’s respond positively to (i.e., the drug reduces plaques), extrapolating that finding to human populations is only justified circularly. Extrapolating the finding to humans requires inferring that the human Alzheimer patients possess some neural pathway (a mechanism, to use Steel’s terminology) that the drug acted on in the mice models. This inference is justified if we can assume that both the target population, humans, and the model organism, mice, are relevantly similar, in that they possess the same mechanism. However,
we do not know that they are relevantly similar. If we knew that both populations were relevantly similar, the extrapolation would not be necessary.

Steel’s extrapolator’s circle can be formalized similarly to Chang’s problem of nomic measurement:

1. Question: Does $S_t$, have M?

2. System, $S$, contains some mechanism, M

3. (1) motivates an extrapolation: If our target system, $S_t$, is relevant in the right way to $S$, we can infer that $S_t$ has M from (1)

4. In order to satisfy the antecedent in (3), we need to know that $S$ and $S_t$ both have M

5. Knowing that $S$ and $S_t$ both have M contradicts (1), eliminating the motivation for (3)

The solution Steel offers to the extrapolator’s circle is a reasoning method he calls comparative process tracing. In comparative process tracing, scientists lack the strong evidence that would confirm the antecedent of (4) above; thus the extrapolation is not self-defeating. What the scientists know instead, is evidence of the underlying causal mechanism that enables them to compare the relevant process in both systems. The regress is broken by having enough knowledge to compare both systems and potentially justify an extrapolation. However, the knowledge is necessarily less than complete, as complete knowledge that the populations are identical would eliminate the need for extrapolation. This ‘sweet spot’ of knowledge requires knowledge about crucial nodes in the respective causal mechanism where the two systems are most likely to differ.

Extrapolation from model organisms to other populations is clearly problematic as we expect there to be differences between species. Extrapolation in psychology is more frequently from
a sample population of human subjects to the general population, also of humans. We might think then that extrapolations from test subjects to the general population are justified.

However, Wissler’s study demonstrates how such extrapolation can go awry. Two (healthy) individuals of the same species may possess roughly the same psychological mechanisms; in fact, much of cognitive psychology takes this to be such a foundational assumption that it is infrequently discussed. Even in cases in which the underlying mechanism may be shared across both sample and target population, the sample population may differ in such a way that the mechanism operates differently. This is the most plausible explanation for Wissler’s failure to find any correlation between reaction times and academic performance. It is unlikely that undergrads at Columbia at the turn of the century had some unique cognitive mechanism(s) related to their intelligence so that there was no relationship between their reaction times and academic performance. Rather, it is likely that sampling bias leads to systematic differences between his sample and the general population that masked the relationship. That his sample was made up of only college students points to the systematic bias being a floor effect on their intelligence. In other words, the range of intelligence in Wissler’s sample was much narrower, and biased towards the top end of the range of intelligence, than the human population at large.

For extrapolation in psychology, the pervasive issue may not be that the mechanisms themselves differ between sample population and target population, but instead that there are individual differences in features between the sample population and target population such that the mechanisms operate differently in a relevant way. Whether psychological populations differ in mechanisms or features is a demarcation problem outside the bounds of this chapter. Regardless of how we demarcate mechanisms, the extrapolator’s regress problem raised by Steel is an additional hurdle for measurement validation.
The extrapolation worry applies directly to our evaluation of external evidence for the validity of a psychological measure. Whereas Stegenga’s concerns about robustness affect a researcher’s ability to determine the independence of two measurement methods, Steel’s concerns about extrapolation affect researchers’ ability to determine the representativeness of the sample that they measure with the measurement method(s) in question. Thus even if a researcher can conclusively establish that, a. the external evidence is concordant, and b. the external measurement measure is independent, there is a further requirement. They also need to establish that their sample, for tests of both measures, is representative of the larger population about which they want to generalize.

**Mechanistic Knowledge in Psychology**

Establishing the representatively of samples requires, on Steel’s account, knowledge of the underlying mechanisms such that we know the sample does not differ in a relevant way to the general population. Once again, there are good reasons to believe that many psychological theories cannot provide this kind of knowledge.

Psychological constructs are, to borrow a term from Feest (m.s. 2017), “lumpy”. These constructs are “lumpy” in so far as multiple different competing definitions make some overlapping claims, but also make very different claims about both the nature of the construct and its potential cognitive mechanism. Feest’s treatment of implicit attitudes exemplifies this lumpiness. Social psychologists have a hard time agreeing about what implicit attitudes are. What they disagree about is the nature of the underlying causal process that creates an implicit attitude. Which is precisely the thing psychologists would need to agree on in order to gain the type of knowledge Steel’s process tracing method of justifying extrapolation requires.
Nosek and Banaji (2009) define an attitude as “an association between a concept and an evaluation” (p. 84). Explicit attitudes are deliberate, intentional, and introspectively accessible associations, while implicit attitudes are associations that are not deliberate, intentional, introspectively accessible. Under their account, it is possible for an individual to have two distinct and conflicting attitudes about a concept, one implicit and one explicit. Gawronski and Bodenhausen’s (2006) Associative-Propositional Evaluation model (APE) distinguishes between implicit and explicit attitudes by associating them with different underlying processes. Explicit attitudes are the result of a propositional process, while implicit attitudes are the result of an associative process. These two processes, propositional and associative, are related to the more general dual-process approaches in psychology (e.g., Chaiken & Trope, 1999). Gawronski and Bodenhausen (2006) claim “people have some degree of conscious access to their automatic affective reactions” while Nosek and Banaji (2009) claim people do not (p. 696).

Both of these accounts include an implicit learning process that is not propositional. Houwer (2014), however, maintains that implicit attitudes are retrieved automatically (and thus are implicit) but are learned via a formation of propositions, just like explicit attitudes. Fazio and Olson (2014a) disagree even further. Their Motivation and Opportunity as Determinants model (MODE-model) assumes, like the APE model, that explicit and implicit attitudes are retrieved via two different processes, but they reject the distinction between implicit and explicit attitudes. They claim both are the same type of mental content (attitudes, i.e., associations between concepts and evaluations) but that they are retrieved in two different ways.

We should worry further about the relative instability of causal mechanisms underlying a construct within and across research projects. Sullivan (2016) argues that constructs in cognitive research are surprisingly unstable. Due to changes in the parameters of an
experimental protocol, we are unjustified in assuming, even when the effects observed are the same, that the causal process leading to the observed effects are the same. Every experiment, even when done by the same researcher in the same lab, will differ, be it in how many times a stimulus is presented, or in the amount of time between a pre-training and testing phase of a task. These differences are exactly the kinds of differences that can affect, or change, the underlying causal process from one experiment to the next. Therefore, it is always an open possibility that two measurement procedures produced concordant data via two different causal processes. She concludes that construct stability:

\[
\text{cannot be met because the terms designating cognitive capacities in cognitive psychology, and particularly in neuroscience, do not have stable referents, and experimental practice in these areas of science currently is not directed at securing such stability. (p. 666)}
\]

Differences between experimental setups include potential differences in the sample being measured. Systematic differences between samples are precisely what caused the failure of early psychologists to find a relationship between RT and intelligence. The worry goes further than just masking relationships between variables, however. As Sullivan points out, even when the effects are the same, i.e., the evidence is concordant, two different causal processes can produce the same effects.

In summary, Steel’s work on extrapolation describes another form of circular reasoning involved in measurement validation. When generalizing findings from a sample population to a target population, researchers must assume that the causal mechanism in the sample population is similar to the target population in the relevant way. Wissler’s study exemplified how this assumption can lead to Type II error, in which a relationship between two variables that exists in the general population is not observed in the experimental sample. The
consequence for measurement validation is that judging the validity of a measure requires extrapolations from sample populations. Wissler failed to observe a correlation between reaction times and academic achievement in his sample population. He, and other intelligence researchers, generalized from this finding to conclude that no systematic relationship existed between reaction times and intelligence, such that reactions times were not a valid measure of intelligence.

The worry for psychology is that the reasoning required to escape from the extrapolator’s circle and prevent errors like Wissler’s requires knowledge of causal mechanisms that we frequently do not have. As with robustness, comparative process tracing sets criteria that many psychological research projects cannot meet.

1.4 Conclusion

Measurement validation in science is difficult, but psychological measures, in particular, seems to be plagued with ongoing problems. Measurement validation exemplifies a form of circular reasoning discussed in the philosophy of science: data-technique circles (Culp, 1995; Collins, 1985). Data-technique circles are any scientific reasoning about experimental data in which one cannot make inferences about the data without presupposing parts of the theory one is testing. The core epistemic problem with measurement validation is the coordination problem: the reliability of a measurement method, and the legitimacy of the construct a measurement method measures, circularly presuppose each other. In order to validate one’s measures, it seems that one must presuppose the very theory that they intend to test. More specifically, a researcher must assume that a relationship exists between the directly-measured variable and the inferred-variable, in order to use her measure to test the relationship between these two variables.
I have argued that the reasons why psychological measurement is especially difficult are due to features of psychological research that impede reasoning about the types of evidence that are relevant to a measure’s validity and can break circular reasoning. A researcher is incapable of escaping this circular reasoning from the boundaries of her individual experiment. However, if she looks outside of her particular experimental design, to other measurement methods, she can find external evidence that can help justify her measurement method. In seeking out external evidence, the research psychologist faces a new set of reasoning problems. Further, due to the nature of psychology, she is hindered in both her ability to create this evidence and evaluate it effectively.

I have argued that in three general features of psychology either impede our ability to evaluate external evidence or make external evidence rare: 1. Even when multiple measures exist of the same phenomenon, the measures are often not independent or fail to produce concordant data 2. Constructs are not stable across labs, or subfields, and 3. Mechanistic knowledge of the phenomenon is absent. When we consider these issues, it is no longer surprising that psychologists have struggled to validate their measures.

I have provided an answer to why psychological measurement is difficult, but that does not mean it is impossible. The problems I have explored in this chapter motivate the need for some method(s) of measurement validation that can accommodate the specific features of psychological research that make the coordination problem seem insurmountable. The rest of this project aims to construct such an account.
Chapter 2

Psychological Construct Validity

2.1 Introduction

Creating valid psychological measures is difficult. Even measures that are well established and widely used are subject to criticism, and sometimes drastic revision. For example, despite being the most commonly used—and one of the oldest—measures of intelligence, some psychologists still object to the Wechsler Adult Intelligence Scale-III based on evidence that it is culturally biased (L. Gottfredson & Saklofske, 2009). What is at stake is the construct validity of the measure. A measure is said to have construct validity when it measures what it is intended to measure, a psychological construct like intelligence. Since the publication of Meehl and Cronbach’s (1955) seminal paper “Construct Validity in Psychological Tests”, construct validity has been both an essential concept in psychological measurement and the source of much confusion and debate. In this paper, I propose a prescriptive account of

\footnote{“Construct” can be replaced with “theoretical term”, which used more commonly in philosophical contexts. I use the term “construct” to respect the standard nomenclature in research psychology.}
construct validity that distinguishes construct validity, a feature of a measure relative to a construct, from construct legitimacy, a feature of the construct itself.\textsuperscript{3}

This account can resolve the debate among psychologists about what construct validity is, and what the requirements are for a test to have construct validity.

In the first section of this paper, I further motivate the need for an account of construct validity by detailing both sources of confusion and disagreement. I argue that psychologists primarily disagree first about whether or not we should distinguish between the adequacy of measures, as measures of a construct, and the adequacy of constructs, relative to their theory. Second, psychologists who do accept this distinction disagree about the relationship between the two forms of adequacy. In the second section, I outline my account of construct validity, and distinguish it from a related, and often conflated, concept: construct legitimacy. In the third section, I demonstrate how failure to distinguish between construct validity and legitimacy leads to premature rejection of useful measurement methods. In the fourth section, I further defend my interpretation of construct validity as an epistemic, rather than ontological, property.

2.2 An Amorphous Concept

Most psychologists would agree that establishing the construct validity of a measure is an essential component of empirical research, even though what “construct validity” means, and what construct validation requires, is ambiguous. Take the following quote from a psychology textbook, for example:

\textsuperscript{3}In the context of psychological research methods, “validity” refers to the property of being sufficiently supported by evidence and does not refer to a property of formalized arguments as it does in logic. For this dissertation, “validity” will always be used in the psychological sense and not the logic sense.
[construct validity] is now generally recognized as the central concern in psychological measurement. Yet construct validity continues to strike many of us, from graduate students to senior professors, as a rather nebulous or “amorphous” concept. (John & Soto, n.d., p. 475)

At first glance, the definition provided in the previous section—that a measure is constructed valid relative to a particular construct when it, in fact, measures that construct—appears adequate. One may wonder, then, where the confusion about construct validity comes from. The ongoing ambiguity about construct validity is due to three related issues. First, explicit definitions of construct validity have evolved over time. Yet psychologists still refer back to Meehl and Cronbach’s seminal 1955 paper when discussing or even defining construct validity. Surprisingly, in this paper, Cronbach and Meehl do not ever explicitly define the term. Rather, they develop an account of construct validation. Second, there is disagreement between these definitions about whether construct validity is an ontological property, of the existence of a particular type of relationship between a measure and a psychological entity, or an epistemic property about one’s belief in such a relationship. Further confusion is caused by psychologists who are sloppy in distinguishing between constructs, which are theoretical terms, and psychological entities, which are the objects that constructs are intended to refer to (see Slaney & Garcia, 2015). It is worth stating that construct validation is essentially an epistemic process of belief justification, i.e. justification of one’s belief that a particular measure is construct valid, a point I do not take to be debatable. What is debatable is whether construct validity is an ontological or epistemic property. Last but not least, there is a further conflation between construct validity as a property of the relationship between a measure and its construct, and construct validity as a property of the construct itself.

In this section, I will explain each of these issues in depth, and show how much of the current confusion and disagreement about how to define construct validity can be traced back the ambiguity in Cronbach and Meehl’s account. Further, I will argue that disagreement between
more recent definitions of construct validity is due to deeper philosophical disagreement about
the relationship between ontological and epistemic claims. Whether one defines construct
validity as an epistemic or an ontological property is determined by how one distinguishes
between the adequacy of a measure and the adequacy of a construct.

2.2.1 Cronbach and Meehl (1955): “Construct Validity in Psychological Tests”

The starting point for any discussion of construct validity is always Meehl and Cronbach’s
(1955) paper, “Construct Validity in Psychological Tests”. In their account, Cronbach and
Meehl (1955) define construct validity as one of four types of test validity. They claim
that construct validity “must be investigated whenever no criterion or universe of content
is accepted as entirely adequate to define the quality to be measured” (1955, p. 282). This
statement of the conditions under which construct validity is relevant is the closest thing
to a definition that Cronbach and Meehl provide. The definition of construct validity often
attributed to them, that a measure is construct valid when it measures what it is intended
to measure, is, in fact, more accurately attributed to R. B. Cattell (1946) or Kelley (1927)
(Borsboom, Mellenbergh, & van Heerden, 2004).

While Cronbach and Meehl do not explicitly define construct validity, they do give an account
of construct validation that rejects operationalist views about psychological constructs. A
good representative of the operationalist view is Anastasi, who claimed, “It is only as a
measure of a specifically defined criterion that a test can be objectively validated at all. .
. To claim that a test measures anything over and above its criterion is pure speculation”
(1950, p. 67). They drew inspiration from the work of logical empiricists in philosophy of
science, in particular, Carl Hempel’s writing on the structure of scientific theories and the
formation of concepts in science (1952). Cronbach and Meehl took Hempel’s nomological network to offer up a way of defining theoretical terms that could not be exhaustively defined by their operations (or associated criterion).

What Cronbach and Meehl refer to as a “nomological network” is an “interlocking system of laws which constitute a theory” (p. 290). Laws, on their account, are generalizations that relate observable properties or quantities, and theoretical constructs to each other. For example, a law of personality traits might describe a relationship between intelligence, a theoretical construct, and G.P.A., an observable quantity. Alternatively, a law could relate two theoretical constructs, like intelligence and working memory. Predictions can be derived from laws that relate theoretical constructs to observable quantities, i.e., criteria.

By utilizing this notion of nomological networks, Cronbach and Meehl explain how a psychological construct is meaningful even if it cannot be exhaustively operationally defined. A construct is meaningful as long as theoretical connections to other constructs exist and it is partially reducible to observables. By partially reducible, they mean that a construct is defined not purely by its operations, but also by its place in the theory: “the construct is not ‘reduced’ to the observations but only combined with other constructs in the net to make predictions about observables” (p. 290).

Cronbach and Meehl’s account of construct validation follows directly from their interpretation of the nomological network. Measures of a construct are validated by testing the hypothesis entailed by the theoretical definition of the construct in question. When a measure produces data that confirms a hypothesis, this is positive evidence for its construct validity. An implicit definition of construct validity can be inferred from their method: A measure is construct valid relative to a construct when the data produced by the measure confirms hypotheses about the construct.
While Cronbach and Meehl’s work is undeniably vital for psychological measurement, the issue of equivocating between construct validity as a property of a measure, and construct validity as a property of the construct is rooted in their account. Those following Cronbach and Meehl have equivocated because their method of construct validation utilizes the same evidence for both construct validation and theory testing. What they describe is hypothesis testing; specifically, testing hypotheses that are entailed by the theoretical definition of the construct-to-be-measured with the measure-to-be-validated. The way one justifies beliefs about a construct, i.e., a particular theory or model of the psychological entity the construct refers to, is by empirically testing hypotheses. If the justification of a belief in a property of a measure utilizes the same evidence, it is plausible to think that the belief that a measure is construct valid is, at least in part, a belief about the theory of the construct.

2.2.2 Campbell and Fiske (1959): “Convergent and discriminant validation by the multi-trait multi-method matrix”

Psychologists following Cronbach and Meehl recognized that their definition of construct validity was ambiguous. In response, D. T. Campbell and Fiske (1959) distinguish between “the adequacy of tests as measures of a construct” and “the adequacy of a construct as determined by the confirmation of theoretically predicted associations with measures of other constructs” (p. 83). Their focus is on the former, rather than the later, and in making this distinction, they claim that Meehl and Cronbach’s notion of construct validity encompasses both types of adequacy.

Campbell and Fiske (1959) advanced the first direct statistical measure of construct validity. They argue that convergent and discriminant validity are two essential types of evidence for the construct validity of a measure. Convergent validity refers to the degree to which
different measures of the same construct agree. Divergent validity refers to the degree to which different measures of different constructs disagree. Agreement and disagreement is determined by the degree of correlation between scores on different measures. Their multi-trait multi-method matrix (3M matrix) is a way of comparing correlations between different measurement methods of different traits (i.e., constructs) so that one can assess the convergent and discriminant validity of a measure. In the 3M matrix, higher correlations between measures of the same construct compared to measures of different constructs are evidence for convergent and discriminant validity. For example, two different intelligence tests should have a higher correlation with each other than an intelligence test and a reading apprehension test if it is the case that the intelligence tests measure a construct distinct from the construct that the reading apprehension test measures.

The introduction of convergent and divergent validity leads to a slightly modified definition of construct validity. For D. T. Campbell and Fiske, a measure has construct validity relative to a construct when the measure has convergent validity and discriminant validity. This definition essentially specifies two specific hypotheses that a measure must confirm in order to be construct valid: 1. Do scores from the measure correlate significantly with scores from other measures of the same construct when taken in similar contexts? 2. Do scores from the measure correlates to a lesser degree with scores from measures of different constructs?.

2.2.3 Messick (1989): Validity

The 3M matrix remains to this day the most common statistical measure of construct validity, and many, if not most, psychologists would endorse some version of construct validation via hypothesis-testing (Smith, 2005). While there has been little methodological advancement about construct validation, there has been a shift in how psychologists conceptualize construct validity. This shift is most evident in the work of Samuel Messick.
Messick’s account of construct validity utilizes Campbell and Fiske’s distinction between two types of adequacy by defining construct validity as the “adequacy and appropriateness of inferences and actions based on test score” (1989, p. 1). This definition limits construct validity to the adequacy of a test as a measure of a construct. Messick’s definition of validity echoes Cronbach and Meehl’s account of construct validation via hypothesis-testing. To validate an inference from a measure “means to ascertain the degree to which multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well supported” (1989, p. 1-2). By understanding validity as a property of an inference, such that a valid inference is a well-justified inference, Messick makes more explicit that construct validity is epistemic. Like Cronbach and Meehl, Messick takes empirical tests of hypotheses about the construct with the measure to be validated as an important form of evidence for the measure’s validity.

Another feature of validity that Messick made explicit is that validity comes in degrees (p. 1-2). This feature is in line with Cronbach and Meehl’s account. They denied that construct validity was an all-or-nothing feature that could be determined with a single statistical measure (1955, p. 290). Messick emphasizes that validity is a graded property in order to support his prescriptive claim that construct validation is a never ending ongoing process. This claim is a direct consequence of a hypothesis testing method of construct validation. As evidence for a measure’s construct validity is produced by empirical testing hypotheses using the measure, any empirical application of a measure provides new evidence for that measure’s validity.

While Messick followed Cronbach and Meehl’s account by fleshing out key aspects of their original definition, he departs from them on two key points. First, Messick collapsed the four types of test validity that Cronbach and Meehl defined into one. Construct validity, rather than being one type of validity, is the ultimate form of validity that subsumes the other types
of test-validity as evidence (p. 7). Second, Messick argued that pragmatic considerations
should be weighed as evidence for a measure’s validity. How a test is used outside of research
applications, e.g., diagnostically in special education programs or as screening for jobs,
requires a form of inference from test scores (p. 8). Therefore, construct validity is a property
not just of inferences from test scores to claims about the construct, but also of inferences
to claims about the individual. Moreover, the justification of inferences to claims about the
individual can utilize pragmatic, moral, or normative reasons. For example using intelligence
tests to place students in school programs results in a higher proportion of African American
students being placed in special education programs, and a lower proportion being placed in
gifted programs. Messick would argue that this consequence of intelligence testing, along
with the moral belief that the resulting inequality is wrong and harmful to African American
students, is a relevant reason for denying that intelligence tests are construct valid measures
for the purpose of school placement.

Messick’s account is a synthesis of the epistemic views of construct validity. Messick, along
with Cronbach and Meehl, Campbell Fiske, and numerous other psychologists, define construct
validity as an epistemic feature of the relationship between measure and construct. According
to epistemic views of construct validity, a measure is construct valid when we are justified in
believing the measure the right sort of relationship to the target construct. They share the
assumption that for a measure to be construct valid it requires having sufficient evidence for
a belief about the measure, but they differ in what counts as evidence and what the content
of the belief is. Messick offers the broadest interpretation of both so that any empirical or
diagnostic application of a test provides new evidence for construct validity, and any inference
based on test scores, not just inferences to claims about a construct, require construct validity.
Cronbach and Meehl have a narrower scope of evidence for construct validity, taking only
on empirical applications of a measurement method that directly test hypotheses about
the construct-to-be-measured. They take construct validity to be a claim both about a measurement method and the measured construct. Finally, Campbell and Fiske take the stricter view that only specific hypotheses are relevant to construct validity, the hypothesis that measures of the same construct should converge with each other, and the hypothesis that measures of different constructs should diverge. Further, they specify construct validity is only a claim about a measurement method as a measure of the construct, not a claim about the construct itself.

2.2.4 Borsboom et al. (2004): “The Concept of Validity”

The final account of construct validity I will describe here radically departs from the previous epistemic views. Borsboom (Borsboom et al., 2004; Borsboom, Cramer, Kievit, Scholten, & Franić, 2009) argues for a revised definition of construct validity that reconceptualizes is as a fundamentally ontological property, rather than an epistemic one, in order to better capture what he takes to be the shared feature of all measures and avoid the absurd consequences of epistemic interpretations. For Borsboom, “When claiming that a test is valid, one is taking the ontological position that the attribute being measured exists and affects the outcome of the measurement procedure.” (2004, p. 1063).

Borsboom et al. (2004) reject the tendency of previous psychologists that started with Cronbach and Meehl, to think of construct validity claims as an epistemic claims about whether or not a particular belief is justified. They argue that we need to return to a definition more aligned with Kelley’s (1927), where a measure is construct valid if it “measures what it purports to measure” (p. 1061). Borsboom et al. propose the following definition:
A test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variation is the attribute cause variations in the outcomes of the measurement procedure (p. 1061).

According to Borsboom, the primary reason we should adopt an ontological view of construct validity is that the existence of the psychological entity to be measured is a necessary precondition for measurement of that psychological entity:

If something does not exist, then one cannot measure it. If it exists but does not causally produce variations in the outcomes of the measurement procedure, then one is either measuring nothing at all or something different altogether. (p. 1061)

For Borsboom, the ontological truth warrants epistemic access (p. 1061). Thus, we need only to worry about determining the truth of the two ontological facts, and the epistemic properties that Messick and others are focused on will inevitably follow. The epistemic views of construct validity have incorrectly reversed the order of this reasoning (p. 1063).

Borsboom is not claiming that the ontological claim and the epistemic claim are the same. He recognizes that they are conceptually distinct (p. 1062). But, “in the case of measurement, it would seem that to talk about the ontology is to talk about the epistemology, and there is surely a sense in which this is correct” (p. 1062).

To further clarify, the epistemic access Borsboom is concerned with here is not epistemic access to evidence that a measurement measures a particular construct. Rather, the epistemic access he refers to is the epistemic access a measurement method affords on a construct (p. 1062). Thus Borsboom is arguing that construct validity is an ontological criterion for an epistemic activity. Even though Borsboom takes measurement to be an epistemic activity, the ontological claims (that the entity being measured exists, and a causal relationship
exists between that entity and the measurement method) are the only universal features of measurement:

The only thing that all measurement procedures have in common is either the implicit or explicit assumption that there is an attribute out there that, somewhere in the long and complicated chain of events leading up to the measurement outcome, is playing a causal role in determining what values the measurements will take. (p. 1063)

According to Borsboom et al. (2004), the epistemic view not only misunderstands what is common to all measurement but also leads to unnecessarily complicated accounts of construct validity. The epistemic characteristics of measurement are superficial and complex. By interpreting construct validity as an epistemic claim, psychologists have mistakenly focused on characteristics that are irrelevant to validity, and created increasingly more complicated accounts that attempt to unify this multitude of unrelated characteristics (p. 1063).

Borsboom et al. also argue that the epistemic view of construct validity leads to absurd consequences. The most damning is that it is possible for a test to be a valid measure of a construct, even if the construct refers to an entity that does not exist (2004, p. 1063-64; 2009, p. 141). Borsboom et al. give the example of phlogiston measures in the 17th and 18th centuries as a case in which the measures of phlogiston would be considered to be construct valid in the past, even though phlogiston, according to current scientific consensus, does not exist (2009, p. 141-142).

Phlogiston was believed to be a type of matter, “fire-stuff”, contained in every substance and emitted when a substance was burned (Bowler, 2005). Phlogiston could be measured by subtracting the weight of a substance after burning from its pre-burning weight. The theory that developed around phlogiston used it to explain a variety of combustion phenomena. For
example, the reason why some materials are more flammable than others is that they contain more phlogiston.

Phlogiston theory was eventually abandoned due to failed predictions and the development of an alternative theory. For example, some materials gain weight when burned. An alternative theory of combustion accurately predicted this phenomenon when phlogiston theory could not. Therefore phlogiston theory was replaced with combustion theory.

Borsboom points out that when phlogiston theory was widely accepted, the measurement methods used to measure phlogiston were construct valid according to any epistemic view of construct validity. There was sufficient evidence, at least initially, that supported various hypotheses about phlogiston. Moreover, scientists developed multiple measurement methods that enabled them to measure phlogiston and produce further evidence that supported phlogiston theory.

Yet, we now believe that phlogiston does not exist. Thus we have a problematic case in which a measurement method was temporarily construct valid. For Borsboom, it is absurd to endorse a theory of construct validity that validates, even if temporarily, measures of non-existent entities. He argues that the phlogiston measures never actually measured phlogiston, and thus we should reject any theory of construct validity that would tell us otherwise.

In summation, Borsboom’s account accepts the distinction made by Campbell and Fiske (1959) between two types of adequacy: 1. the adequacy of a test as a measure of a construct, and 2. the adequacy of a construct. However, Borsboom restricts construct validity to the latter adequacy. The primary question of construct validity is then not a question of whether a measure measures a construct, but rather is a question of whether the construct refers to a psychological entity that exists, and whether a causal relationship exists between that entity and the measurement method.
Even if one agrees with Borsboom’s ontological view, a question remains about how it is applied to the practice of construct validation. Construct validation is an undeniably epistemic process, in which one justifies a belief in the construct validity of a measure. Even if we take the ontological view of construct validity, we have only changed the content of the belief. The belief still needs to be justified, and we may think that the justification process is the same regardless of whether construct validity is an ontological or epistemic property.

However, psychologists interpreting Borsboom’s view and applying it to their own measures seem to believe that it does involve a radical revision of construct validation. Take De Houwer et al.’s (2009) evaluation of implicit measures using Borsboom’s criteria. De Houwer et al. (2009) take the second criteria, that variation in the psychological entity causes variation in the test, to require specifying the causal mechanism underlying the psychological entity that explains variation in test scores. Therefore, validating the Implicit Association Test (IAT) requires knowing the process through which changes in the strength of implicit associations cause changes in an individual’s reaction times. They acknowledge that this criterion is not currently met by the IAT or any other implicit measure for that matter:

> For most measures it is not entirely clear what they measure, what processes produce the measure, and whether the processes are automatic in a certain manner. (De Houwer et al. 2009, p. 262)

Yet even though most tests fail to meet their criteria for construct validity, they still believe that these tests are useful research tools (p. 262). It seems then that for De Houwer et al., implementing Borsboom’s definition of construct validity creates a set of criteria that are ideals for psychologists to strive towards rather than preconditions for the use of a measure in empirical contexts.
De Houwer et al.’s implementation of Borsboom’s view also demonstrates that an ontological view of construct validity ends up making the adequacy of a construct (the latter form of adequacy from Campbell Fiske) a necessary condition for construct validity. Recall that Campbell and Fiske refer to this adequacy as the adequacy of a construct with respect to the theory it is a part of, and the evidence for that theory (1959, p. 83). Determining the causal mechanism that links variation in a psychological entity and a measure seems like it is directly relevant to the adequacy of the construct, which refers to that psychological entity and its underlying causal mechanism. Thus Borsboom’s account of construct validity takes the adequacy of a construct to be a necessary condition for the validity of a measure.

### 2.2.5 An Amorphous Concept, Revisited

In this brief history of construct validity, I have described five distinct definitions of the term. According to these different accounts, a measurement method, m, of a construct, C, is construct valid iff:

1. m, in fact, measures C (Standard-View) (R. B. Cattell, 1946; Kelley, 1927)

2. m confirms hypothesis entailed by C’s theoretical definition (Cronbach & Meehl, 1955)

3. m has convergent validity with other measures of C, and divergent validity with measures of constructs that are not C (D. T. Campbell & Fiske, 1959)

4. our inferences and actions based on m are sufficiently justified (Messick, 1987)

5. (a) the psychological entity that C refers to exists, and (b) a causal relationship between m and C exists such that any variation in C will cause variation in m (Borsboom et al., 2004, 2009)
I have argued that the majority of these definitions (2-4) are variations of an epistemic view of construct validity. Each definition takes construct validity to be about whether or not we are sufficiently justified in believing that a measurement method measures the intended construct. Where these accounts differ is primarily in what they take to be relevant evidence for construct validity. In contrast, ontological views of construct validity (1 and 5) take construct validity to be a property of the psychological entity to which the construct refers, and that the measure tracks. What De Houwer et al.’s examination of implicit measures demonstrates is that the disagreement between epistemic and ontological views of construct validity is more than just a philosophical debate about defining a concept. To take an ontological view of construct validity means adopting criteria that most (if not all) psychological measures fail to meet.

Despite the disagreement and confusion surrounding construct validity, it is an important and useful concept used both in ongoing debates about new and old psychological measures, and current philosophical projects in philosophy of cognitive science. In personality psychology, for example, construct validation of personality tests is ongoing.; a search for articles concerning the construct validity of personality measures shows more than 1,000 articles were published on the topic in 2017 alone (Google Scholar, 2017, January 17). In philosophy of cognitive science, philosophers are concerned about construct validity in specific cases, for example, Feest’s work on the construct validity of the Implicit Association Test (m.s. 2017), along with more generally how construct validity affects efforts to localize cognitive capacities in the brain (Sullivan, 2016; Poldrack & Yarkoni, 2016; Poldrack, 2006).

In the next section, I will explain my account of construct validity that seeks to provide much-needed clarification about what construct validity is and what the requirements for establishing it are. My account provides an epistemic definition of construct validity that maintains Campbell and Fiske’s (1959) distinction between the two forms of adequacy.
Unlike Borsboom's ontological view, my account explains how current measures used in psychology can be evaluated as construct valid independently of our evaluation of the constructs themselves.

2.3 A Model of Construct Validity

In the previous section, I highlighted two types of adequacy underlying Meehl and Cronbach’s original definition: adequacy of a test as a measure of a construct, and adequacy of a construct. Current disagreement about the criteria for construct validity is due to differing views about the relationship between these two types of adequacy. The account of construct validity I give here maintains this distinction, but I define construct validity as referring only to the former type of adequacy, and not the later. The latter type of adequacy I call construct legitimacy. The difference between construct legitimacy and validity can be illustrated thusly:
Figure 2.1
Construct Validity
The above model shows three relationships between the world, a measure, and a construct. Construct validity is a property of the relationship between a measure and a construct, such that a measure has construct validity relative to a specific construct if we have sufficient evidence that it tracks that construct. To claim a measure has construct validity requires, at bare minimum, that we have evidence we have isolated a distinct construct. The convergent and discriminant validity from Campbell and Fiske’s (1959) account is exactly the type of evidence that would indicate that a measure isolates a distinct construct.

Construct legitimacy, in contrast, is a property of a construct itself, relative to the theory or model it is a part of. A construct can be said to be more or less legitimate depending on how well the definition and/or model of the construct are supported by the available evidence, and how useful the construct is in the larger theory. We say that a construct is legitimate, relative to a particular model or theory when that model or theory is sufficiently justified. Roughly, construct validation is about the justification of inferences from test scores to claims about a construct, while construct legitimization is about the justification of a construct’s theory/model. The criteria for the legitimacy of a construct will depend on both our standards for evaluating theories and our metaphysical commitments about the referents of constructs.

Existence claims about a construct are claims about the relationship between the construct, a theoretical concept, and the object(s) in the real world to which it is intended to pick out. As legitimacy depends on metaphysical commitments about the psychological entities that constructs refer to, a realist about psychological entities may not take construct legitimacy and existence claims to be distinct.

My definition of all three concepts here is purposefully theory-neutral. I make no commitment in any of these definition about what the criteria should be for each. Establishing criteria
requires making epistemic commitments about the justification of these claims, along with
metaphysical commitments about the relationship between constructs and psychological
entities. My purpose here is to merely argue that these two concepts, legitimacy and validity,
come apart such that they can vary independently of one another.

Construct validity applies to specific inferences made on the basis of scores from a measure.
My definition here follows Messick’s by focusing on the justification of inferences from test
scores. However, I take construct validity to apply to only inferences to claims about a
construct, rather than any inference. My definition can be formalized as follows:

A measure, \( m \), of some construct, \( C \), is construct valid iff we are sufficiently
justified in believing \( m \) tracks \( C \).

A measurement method can have construct validity relative to one construct, but not relative
to another. This leads to claims of the following sort: “The SAT is a construct valid measure
of school performance, but not of general intelligence”. More generally, construct validity
claims will follow this phrasing, “X is a construct valid measure of construct Y,” or, “X has
construct validity relative to Y”.

Construct legitimacy, as a property of constructs themselves, applies to the particular
definition and/or model of a construct. Psychological constructs are theoretical concepts that
are part of a larger theory. The legitimacy of a construct is determined by how complete
the model/definition of a construct is, how well the definition is supported by the empirical
evidence, and how integral the construct is to the larger theory. While I disagree with De
Houwer et al.’s (2009) criteria as it pertains to construct validity, explicating the causal
mechanism underlying a construct provides strong evidence for construct legitimacy. As
legitimacy is dependent on the status of the construct’s theory, legitimacy claims will often
appeal to the particular theory. For example, “Phlogiston is an illegitimate construct because phlogiston theory is not supported by evidence”.

In distinguishing between construct validity and construct legitimacy, I allow both features to vary independently of one another. This does not mean that the two are not partially dependent on one another. Clearly whether or not a particular test measures an intended construct will depend on what we think that construct is. However, my distinction here puts me in direct contrast with De Houwer, Teige-Mocigemba, Spruyt, and Moors (2009) and Borsboom et al. (2004), who take construct legitimacy as a necessary prerequisite for construct validity. They set an incredibly high standard for psychological constructs such that most psychological tests do not meet the standards for construct validity. I propose instead that if construct validity is to be a useful concept for psychology, we should take it to be a property that can be limited, but not prohibited, by the degree of legitimacy of the target construct. Thus the construct validity of a measure requires that the construct it purportedly measures adequately captures a psychological object. However adequately capturing a psychological object does not necessitate that we have a complete model or explanation of that psychological object, or that the object truly exists.

The final part of my account is that construct validity (and construct legitimacy) comes in degrees. A test is more or less valid relative to a particular construct depending on the evidence for the relationship between the test and the target construct. Likewise, a construct is more or less legitimate depending on the evidence for the associated theory, or, as we will see in the next section, depending on whether there are competing theories of the same construct. While not present in the original definition by Meehl and Cronbach, this view is shared by both psychologists and philosophers writing about construct validity more recently (Smith, 2005; Messick, 1995a, 1995b; Feest, 2017). A natural consequence of this view is that
both validity and legitimacy are evolving properties, such that a construct or measure may become be more or less legitimate/valid over time as new evidence is gathered.

Whereas Borsboom et al. (2004) require establishing both a causal relationship and the existence of a psychological entity in order to claim a measure has construct validity, I require only that the measure isolates a distinct construct. By isolating a distinct construct, I have in mind Campbell and Fiske’s (1959) discriminant and convergent validity. A measure then isolates a distinct construct if it converges with other measures of the same construct, but discriminates from measures of distinct constructs.

There are good a priori reasons to distinguish between construct validity and legitimacy. It is easy to find examples of cases in which a test irrefutably has construct validity, even though the construct itself is an illegitimate one based on any scientific standard. Take any test of your zodiac sign. Your zodiac sign depends on your birth date, something that we are incredibly accurate at measuring. Any zodiac sign test that accurately determines your birth date, say by asking you to self-report, or even looking up your birth date via public records, will correctly measure your zodiac sign. However, no psychologist (that I know of) would claim zodiac signs are a legitimate psychological construct, that are able to explain or predict any variation between individual’s personality traits.

In this case, there is sufficient evidence for the construct validity of zodiac sign tests even though zodiac signs are illegitimate. The test is highly accurate and reliable, converges with other tests of zodiac signs, and discriminates from measures of different constructs. Crucially, the test isolates a distinct construct. There is no question that what the test reports back is a zodiac sign and not some other astrological construct. Here we see that construct validity responds to a variety of evidence that is irrelevant to the legitimacy of the construct in question.
2.4 Two Examples from Research Psychology

I now turn to two debates in psychology that illustrate how construct validity and legitimacy come apart in psychological research. These debates show not only that construct validity and construct legitimacy can vary independently of one another—that is to say, a measure can have reasonable degree of construct validity even when the construct it measures has a low degree of construct legitimacy—but also why such a distinction is necessary. I will argue that these two cases demonstrate a dissociation between construct legitimacy and construct validity, in which measures have construct validity despite the relative illegitimacy of the constructs they intend to measure. Trying to understand either case without distinguishing between construct legitimacy and validity would lead to erroneous abandonment of measures that are useful research tools.

2.4.1 The Implicit Association Test

The Implicit Association Test (IAT) (Greenwald et al., 1998) measures implicit attitudes by comparing reaction times between paired concepts. Participants are instructed to sort stimuli that are examples of four concepts; however, in each trial, the concepts are split into two pairs (ex. Male-career; female-home) so participants sort four types of stimuli into two categories. The pairs are varied across trials, and participants are timed as they sort the stimuli. Researchers infer a subject’s implicit attitudes from a comparison between average reaction times to each association pair. The faster a participant is able to sort stimuli from a paired concept, the stronger that subject’s implicit association between the concepts. Thus, if a participant sorts male-career stimuli faster than female-career stimuli, they would be judged to have a stronger implicit association between males and careers.
Despite the popularity of the IAT, widespread disagreement remains about what exactly implicit attitudes are. One major source of disagreement is about the process of learning and retrieving implicit attitudes. Nosek and Banaji (2009) define an attitude as “an association between a concept and an evaluation” (p. 84). Explicit attitudes are deliberate, intentional, and introspectively accessible associations, while implicit attitudes are associations that are not deliberate, intentional, introspectively accessible. Thus the difference between implicit and explicit attitudes on their account rests on different processes of retrieval.

Gawronski and Bodenhausen’s (2006) model, on the other hand, take explicit and implicit attitudes to be the result of two different learning processes. Drawing on dual-process theories, they propose that explicit attitudes are the result of a propositional process, while implicit attitudes are the result of an associative process. Gawronski and Bodenhausen claim “people have some degree of conscious access to their automatic affective reactions” while Nosek and Banaji (2009) claim people do not (p. 696).

Fazio and Olson (2014b) agree with Nosek and Banaji that an attitude is implicit or explicit based on how it is retrieved. However, they reject distinguishing between implicit and explicit attitudes as different types of mental content. Rather, they take both to be the same type of mental content that are retrieved in two different ways.

Based on this theoretical disagreement, Feest (m.s. 2017) concludes that implicit attitude is a “lumpy construct”. By this, she means there is some agreement (there exist implicit and explicit retrieval processes of attitudes that result in different responses) along with widespread disagreement (whether the different processes retrieve different types of representations, what form the representations take, and how the representations are formed).

This type of definitional lumpiness is a common for psychological constructs and detracts from their legitimacy. The theory of implicit attitudes is currently incomplete, as there is
no consensus about many aspects of the construct. The incompleteness of the associated
theory, in turn, limits the legitimacy of the implicit attitudes construct. Low legitimacy does
not necessarily indicate a construct should be abandoned. As Feest (m.s. 2017) points out,
implicit attitudes have an explanatory role, particularly in cases in which explicit attitudes
are unable to predict behavior accurately. Implicit attitudes have this explanatory role even
if the theory is incomplete. Rather, implicit attitude’s low degree of legitimacy reflects the
need for more evidence to arbitrate between the multiple, conflicting hypothesis about the
nature of the construct.

The question relevant to my account is whether the construct validity of the IAT as a measure
of implicit attitudes is prohibited or merely limited by the relatively low legitimacy of the
construct. Some psychologists, most notably De Houwer et al. (2009) have argued that it
cannot because we cannot explicate the causal mechanism underlying implicit attitudes that
would explain scores on the IAT. Feest, in contrast, believes the IAT has construct validity
despite the issues with implicit attitudes.

On my account, the construct validity of the IAT is not wholly dependent on the legitimacy
of the implicit attitudes construct. Rather, the myriad of other forms of evidence for
the construct validity of the IAT can be sufficient despite the target construct’s relative
illegitimacy. Feest points to the explanatory role IAT scores have for certain discriminatory
behavior, even when predictive validity is low. There is also significant convergent evidence
for the IAT: IAT measures correlate strongly with explicit attitude measures and with other
implicit measures. Importantly, the strength of this association (between implicit and explicit
measures of the same attitude) is domain dependent, and mediated by an individual’s
concern with expressing the attitude (Nosek, 2005). This means that implicit attitudes
usually correspond with our explicit attitudes, except in situations where an individual feels
pressure not to express a particular attitude.
Taking this evidence into account, I agree with Feest (m.s. 2017) that the IAT has construct validity relative to implicit attitudes. Results from IAT have predictive power, and psychologists have argued that even when effect sizes are small, the practical implications can be large. However, the degree of construct validity could be improved on in the future if we develop a better, unified, model of implicit attitudes that in turn leads to improvements of the IAT. Note that the positive relationship between legitimacy and validity here is indirect. Increasing the legitimacy of the measured construct does not necessarily increase the validity of the construct’s measures. Increasing legitimacy increases validity only when such theoretical advances enable improvements in the measure itself, for example through identification of previously unknown confounding variables that can now be controlled for.

2.4.2 Implicit Memory

Roediger’s argument against implicit memory demonstrates how a construct can be illegitimate even when the measures of the construct themselves have a significant degree of construct validity. Implicit memory exemplifies a different sort of illegitimacy than implicit attitudes. According to Roediger, implicit memory is an illegitimate construct due to a lack of evidence that the construct refers to a uniquely cognitive phenomena, not because the construct is lumpy. Compared to implicit attitudes, there is a surprising lack of disagreement about how to define implicit memory.

Implicit memory is a psychological construct that is usually defined in contrast with explicit memory. Explicit memory refers to memory that is intentional and therefore can be both recalled and reported by a subject. Implicit memory, in contrast, is unintentionally recalled. Implicit memory is commonly measured using motor skill learning tasks, where subjects learn a physical based skill, like tracing a circle with their finger, or priming tasks, like word-stem and word-fragment completion. In priming tasks, subjects are primed (i.e., presented) with
a set of words, then later tested with both the primed (old) words and new words. Faster reaction times to primed words indicate a priming effect (Roediger III, 2003).

Warrington and Weiskrantz (1970) demonstrated that amnesic patients had impaired explicit memory but did not demonstrate impaired performance on word-stem and word-fragment completion tasks. In order to explain the single dissociation, psychologists proposed a new memory capacity: implicit memory. Amnesic patients have unimpaired implicit memory, but impaired explicit memory, explaining why they scored normally on priming tasks (which measure implicit memory) but not on explicit memory tasks.

While this single-dissociation evidence has been replicated numerous times, Roediger (2003) argues that implicit memory is an illegitimate construct. According to the definition of implicit memory used in the research literature almost any biological system, regardless of whether it involves the brain or not, can be said to have implicit memory (p. 4). The problem with this definition is that any change in response to a repeated stimulus, not just psychological, counts as memory. If this change in response does not involve any intentional recollection, then according to McDermott (2000) and others, it is implicit memory. But if implicit memory is just the manifestation of memory absent of intentional recollection then any biological system, like the immune system, has some sort of implicit memory. When the immune system is attacked with a pathogen, it has previous experience with, its response changes, reflecting a “memory” of the pathogen.

This example shows a disconnect between the broad theoretical definition of implicit memory and its narrow operationalization. The problem here is not simply that psychologists have failed to measure the entire set of phenomena that could be considered cases of implicit memory. Rather, the problem for Roediger is that the range of phenomena covered by the theoretical definition of implicit memory is so broad that it includes phenomena psychologists
would not consider to be cases of implicit memory at all, regardless of their ability to measure them.

He concludes that any commonality we may see between instances of implicit memory only exist at a level of general principle rather than at the level of a specific mechanism. Biological systems respond differently to stimuli they have encountered before, but there is no implicit memory capacity explained by some specific cognitive mechanism (p. 13). Put more simply, Roediger believes that legitimate psychological constructs must have a unique cognitive mechanism.

If the implicit memory construct is illegitimate, it cannot be made legitimate by merely redefining the construct to refer to only cognitive phenomena. Roediger views redefining the construct as ad hoc, as without a unique cognitive mechanism, implicit memory is not a uniquely cognitive phenomenon. Rather, it is “a memory function frequently incidental to the operation of the system which subserves a difference basic function” (p. 14).

It is hard to make sense of Roediger’s argument here unless we distinguish between legitimacy and validity. Roediger has no issue with the actual measures of implicit memory, or the effects observed using these measures. His only issue is with the justification for proposing implicit memory as a distinct memory capacity. The single dissociation evidence, while potentially sufficient for the initial distinction, is no longer sufficient for Roediger in light of the failure to identify a cognitive mechanism. In other words, Roediger believes implicit memory is an illegitimate construct, even though the measures associated with the construct have construct validity. They measure what we intend for them to measure; however, what they measure is not a uniquely psychological phenomena.

If a unique cognitive mechanism was identified, thereby legitimizing the implicit memory construct, this would not likewise increase the construct validity of implicit memory measures,
unless indirectly. One could imagine a cognitive mechanism that narrowed the scope of implicit memory to only perceptual domains. This would then, in turn, eliminate some implicit memory measures, and potentially increase convergence across the remaining measures.

2.5 Further Defense of the Distinction

I have argued that the previous two examples show how construct validity and construct legitimacy can vary independently of one another. I have also argued that understanding these examples requires a graded notion of both construct validity and legitimacy. Both the IAT and priming measures are construct valid, despite the problems with the constructs they measure.

De Houwer et al. (2009) in contrast denies that any measure of implicit attitudes have construct validity because the construct itself is not well understood, e.g., has low construct legitimacy. Similarly, for De Houwer et al., Roediger’s rejection of the construct “implicit memory” would entail denying the tasks used to measure implicit memory have any construct validity. My alternative reading of these cases is more permissive. We can talk about the construct validity of measures in cases where the causal mechanism is unknown if we use a notion of construct legitimacy that varies independently of construct validity. Construct validity requires then a bare minimum of construct legitimacy, in the form of evidence that the measure tracks a distinct construct, rather than the full mechanistic description required by Borsboom.

My argument thus far has been conceptual, focusing on the distinction between adequacy and legitimacy, and not the criteria for construct validity that follow from Borsboom et al.’s (2004, 2009) account. A skeptic may grant this conceptual distinction but still maintain that construct validity requires a high degree of construct legitimacy. I believe there is a further
pragmatic justification for rejecting legitimacy as a necessary prior for validity. If construct validity requires establishing construct legitimacy to the degree that the causal mechanism linking a construct to the results of its measure, construct validity ceases to be a useful concept. Borsboom’s criteria would force us to either accept that most psychological measures are not construct valid or promote “unproductive debates over conceptual identification of attributes underlying psychological measures” (Nosek & Greenwald, 2009, p. 373). Construct validity would be a regulative ideal. We could talk about how a measure is closer to construct validity than another rather than talking about the degree of construct validity a measure has. But absent from such talk is a standard for determining when a measure is appropriate for empirical research. If no measures, or at least very few measures, are construct valid, what justifies their use? When construct validity is understood as a regulative ideal, we create a conceptual void in which we need an additional notion that prescribes which measures are good enough for research. Rather than relegate construct validity to an ideal and create a need for a new standard for test use, I prefer using the concept we already have.

On my account, it is not the case then that the IAT has no construct validity because the construct it measures, implicit attitudes, is an ill-defined construct. Rather, the degree to which the IAT has construct validity is in part lessened, but not prohibited, due to our lack of understanding about the construct it is intended to measure. The construct validity of the IAT can be improved upon by increasing our understanding of implicit attitudes, which in turn would increase the legitimacy of the implicit attitude construct. However, there are other types of evidence that would increase the construct validity of the IAT independent of the legitimacy of the implicit attitudes construct. To name just a few, further increasing the IAT’s predictive power for future behavior, or decreasing variability between scores from the same participant taking the IAT at different times (also known as test-retest reliability), would increase the construct validity of the IAT without increasing the legitimacy of the
construct it measures. That does not mean increasing the construct validity of the IAT would not, in turn, increase the legitimacy of the implicit attitudes construct. However, that improvement would be indirect, by way of more research on implicit attitudes with an improved, more powerful IAT that is subject to fewer confounding variables and sources of score error.

My account here explains current practices in social psychology, where the IAT is an incredibly useful measure. By distinguishing between legitimacy and validity, we can explain how the IAT has sufficient construct validity to be used in psychological research even though the construct it measures is ill-defined. Moreover, using a graded notion of construct validity allows one to claim the IAT has construct validity even though the degree of construct validity it has could be improved.

In addition, if implicit memory is an illegitimate construct, my account allows us to rejecting the construct without rejecting the tasks used to measure it. If we fail to distinguish between legitimacy and validity, we erroneously reject priming tasks due to the problems with justifying implicit memory as a distinct memory capacity. There currently are no psychological mechanisms that can explain how variation in implicit memory causes variation in priming effects, but we must recognize that this as hindering the legitimacy of the construct, not the validity of the measures. Again, my account is able to explain current practices in psychology, where these semantic and visual priming tasks are valid measures in their own right with applications outside of just implicit memory studies.

Borsboom’s account of construct validity does more than turn construct validity into a regulative ideal. He reinterprets construct validity as an ontological, rather than epistemic property. By adopting an ontological view of construct validity, Borsboom avoids two issues. First, it appears that epistemic views have to permit the absurd consequence that
a measure’s construct validity varies over time. With the phlogiston example, Borsboom demonstrates that measures of phlogiston were construct valid according to epistemic views of construct validity, and at some point ceased to be construct valid when phlogiston theory was abandoned. Second, when we interpret construct validity epistemically, we are forced to accept increasingly more complicated accounts of justification that attempt, and fail, to provide cohesive explanations of measurement validation.

Borsboom et al. (2004) correctly identifies these consequences of epistemic accounts of construct validity. However, he is wrong to interpret either as real problems of an epistemic view. In my account, these are features, not bugs. The phlogiston example demonstrates how construct validity is an evolving property that is related to the justification we have for the theory of the construct being measured. It is true that evidence that falsifies a theory can then decrease the construct validity of the related measures. But this is exactly what we should expect for any form of belief justification. Justification does not guarantee truth regardless of the particular belief or epistemic theory.

The difficulty in creating a cohesive explanation of construct validation is also a feature of justification in general, not merely construct validation. There are many different forms of evidence that are relevant to the validity of a measure, and they do not cease to be relevant by redefining construct validity ontologically. Borsboom rightly acknowledges that it is difficult to provide a descriptively accurate account of measurement justification, in part due to the variety of relevant evidence. However, that a task is difficult does not negate its value. Rather, the difficulty here reveals the need to a rigorous philosophical examination of measurement validation. If we accept that construct validity is an ontological, we may neglect the work in epistemology that may help explain construct validation as belief justification.
Borsboom is mistaken when he claims that the truth of the ontological claims about psychological entities and causal relations guarantees epistemic access. Measurement of a construct requires epistemic access that is contingent on human factors. There are examples of this contingent relationship throughout the history of science. Neuroscientists could not measure the BOLD-response prior to the invention of fMRI. Our epistemic access to the phenomenon is contingent on the series of innovations in biomedical imaging that lead to the development of fMRI. What we can measure is determined not merely by what actually exists, but also by the technology of the time and the status of the relevant scientific theories.

Even if one is a staunch realist about the referents of constructs, determining whether or not a measure tracks a construct requires evaluating evidence. This is an epistemic problem, not an ontological one. In other words, even if the ontological claims are true about a measure, scientists need to determine whether they are justified in believing those claims. There is no way to avoid the epistemic problem of construct validation.

### 2.6 Construct Validity and the Coordination Problem

I argued the previous chapter that measurement validation in psychology faces a coordination problem. Establishing that a measurement method measures a particular construct seems to entail a type of viciously circular reasoning in which we are forced to presuppose a relationship between the quantity we can directly measure and the construct we want to infer. Solving the coordination problem requires an epistemic theory of justification that can be applied to construct validation. However, not just any epistemic theory will do.

There are many different epistemic theories of justification. However, a candidate theory needs to do more than provide an account of belief justification to be a successful account of
construct validation. I propose the following three criteria for evaluating a potential epistemic theory of construct validation:

1. That it enables evaluative claims about the construct validity of a particular measurement method.

2. That it solves the coordination problem.

3. That it is applicable to psychological cases.

4. That it respects the distinction between construct validity and construct legitimacy.

The first, and most basic, criterion that a candidate epistemic theory must meet is that it enables evaluative claims about a measurement method’s construct validity. These can be sufficiency claims about the construct validity of an individual measurement method, i.e., “This intelligence test is a construct valid measure of intelligence,” or “The SAT is not a valid measure of intelligence.” They can also be comparative claims between measurement methods, i.e. “this intelligence test is a more valid measure of intelligence than the SAT” or “the revised form of the SAT is a better measure of school performance than the original test.” I have argued here that construct validity is a graded property. Thus, a potential epistemic theory will need to establish what factors increase or decrease construct validity. Ideally, the theory will also tell us what conditions must be met in order for a measure to be sufficiently construct valid for use in research.

The second criterion refers to the coordination problem outlined in the previous chapter. As I have argued, solutions that have been offered for other forms of circular reasoning in science are inadequate here due to the nature of psychological research and psychological constructs. This leads directly to my third and fourth criteria.
The third criterion is that an epistemic theory must be able to deal with the obstacles created by the nature of psychological constructs. While psychological constructs are not different in kind from theoretical terms employed in other sciences, psychological constructs, on the whole, tend to be under-defined, refer to a variety of different types of psychological objects, and play a multitude of explanatory and/or predictive roles depending on the research program.

Constructs like “anxiety”, “working memory”, or “intelligence” have long established research programs, yet are under-defined in the sense that there is no general consensus among psychologists in these research programs about either the correct definition of the construct or about the particular model and/or theory of which the construct is a part. This is the problem that Meehl and Cronbach refer to when they lament the general “underdetermination of psychological constructs” (1955, p. 294). Meehl and Cronbach’s specific concern was the lack of laws in psychological theories that would in part define a construct under a syntactic view of scientific theories. However psychological constructs are under-defined even if one does not require definitions to cite laws. Many constructs are also “lumpy”. These constructs are “lumpy” in so far as there are multiple different competing definitions that may make some overlapping claims, but also make very different claims about both the nature of the construct and its potential cognitive mechanism.

The final criterion is essentially a requirement of descriptive adequacy for a theory. Construct legitimacy and validity must be distinguished between in order to account for construct validation as psychologists currently practice.

### 2.7 Conclusion

Construct validity is an important and useful concept in research psychology when understood correctly. I have argued that criticisms of construct validity have arisen partly due to a failure
to adequately distinguish construct validity from construct legitimacy, and a failure to use a graded notion of both.

Distinguishing between these two concepts clarifies certain debates in psychology about both measures and constructs. I have shown here how my account differentiates between issues of construct validity, where the measure of a construct is the focus (like in the debate about the IAT), and issues of construct legitimacy, where the usefulness of the construct itself is in question (like in Roediger’s argument against implicit memory). By taking both legitimacy and validity as things that come in degrees, my account is able to further explain how measures like the IAT have a degree of construct validity that affords their current use in research while allowing for future improvement as our understanding of the constructs they measure deepens. Moreover, by adopting this distinction between legitimacy and validity, we can avoid erroneously throwing out measures, like priming tasks, that measure problematic constructs, like implicit memory. Finally, by distinguishing between legitimacy and validity, I was able to demonstrate how each responds to different types of evidence. While both construct legitimacy and validity are indirectly related, neither is necessary or sufficient for the other. Thus, a measure’s construct validity may be inhibited by the lack of construct legitimacy for its intended construct, it is not, in principle, prohibited by it.

In addition, the account I have argued for here restricts possible epistemic theories of construct validation. A potential epistemic theory of justification must not only solve the coordination problem but must also respect the distinction between legitimacy and validity. In the following chapters, I will examine three candidate epistemic theories: 1. operationalism, 2. hypothesis-testing, and 3. coherentism.
Chapter 3

Psychological Operationalism

3.1 Introduction:

I argued in the previous two chapters that construct validation in psychology faces a coordination problem. Establishing that a measurement method measures a particular construct involves potentially viciously circular reasoning in which we are forced to presuppose a relationship between the quantity we can directly measure and the construct we want to infer. Solving the coordination problem requires an epistemic theory of justification that can be applied to construct validation. However, not just any epistemic theory will do. As I concluded in the previous chapter, a candidate epistemic theory must:

1. Enable evaluative claims about the construct validity of a particular measurement method.

2. Solve the coordination problem.

3. Be applicable to psychological cases.
4. Respect the distinction between construct validity and construct legitimacy

Satisfying these conditions will require further refinement of the general characterization I gave in the previous chapter of the relationship between measure and construct. In this chapter, I will evaluate the original epistemic theory of measurement validation in psychology, operationalism. The failure of this view is due to how it characterizes the relationship between measure and construct. An analysis of how and why operationalism fails, then, aids in our understanding of how measure and construct relate.

Operationalism is the view that the meaning of a construct is the set operations by which the construct is measured. For operationalists, a construct is synonymous with its measurement method. This view characterizes the relationship between measure and construct as one of direct entailment. Recall that Cronbach and Meehl (1955) introduced construct validity as a counter to the operationalist views that prevailed among psychologists in the first half of the 20th century. In this chapter, I delve deeper into the history of operationalism in psychology in order to investigate its potential as an epistemic theory that could solve the coordination problem.

It may seem strange to begin with a theory that the first proponents of construct validity rejected. However, as I will show, Operationalism as formulated and employed by research psychologists is not necessarily in direct conflict with construct validity. For example, D. T. Campbell and Fiske (1959) are credited with developing the first direct statistical measure of construct validity. Yet they were committed to operationalism. Campbell, in particular, wrote extensively about convergent operationalism (D. T. Campbell, 1960, 1988; Marshall, 1979), a variety of operationalism I will examine in depth in this chapter.

Moreover, versions of an operationalist view are still advocated for by psychologists. For example, (van der Maas et al., 2014) argue that the mutualism model of intelligence, an
alternative to the g-factor model of intelligence, implies that general intelligence is nothing more than a weighted sum score. They claim that if the mutualism model is correct, intelligence is just what intelligence test measure.

As the above discussion indicates, there are multiple versions of Operationalism. There is Operationalism as initially formulated by Percy Bridgman, and then there is Operationalism as interpreted and applied by psychologists. I begin in the first section with Bridgman’s operationalism, henceforth referred to as strict operationalism. Bridgman was heavily influenced by logical positivism. Many failures of strict operationalism trace back to the failures of logical positivism’s verifiability criterion of meaning. However, there are additional reasons why strict operationalism is untenable as an epistemic theory of construct validation in psychology that I will discuss here. In the second section, I analyze methodological operationalism, a version of operationalism (Feest, 2005) attributes to the first adopters of operationalism in research psychology. This view, I argue, in inadequate as an account of measurement validation primarily because it cannot account for construct validation by hypothesis testing. Finally, in the third section, I look at convergent operationalism, a form of operationalism that originated in work by psychologists like Campbell in the 1950s but has been defended more recently by psychologists (C. Green, 1991). I argue that by attempting to avoid the pitfalls of strict operationalism, convergent operationalism ends up endorsing philosophically inconsistent claims about the nature of unobservable psychological constructs. Thus even this form of operationalism fails as an epistemic theory of construct validation.

### 3.2 Strict Operationalism

*Strict operationalism*, as formulated most famously by Percy Bridgman (1882-1961), claims that a construct is defined by its method of measurement alone (1927). For the strict
operationalist, there is no meaning to a theoretical term beyond the established measure of that term. Bridgman writes, “we mean by a concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations.” (1927, p. 5). Bridgman’s notion of operations, e.g., a method of measurement, is highly restrictive in that there is only one measurement method per construct:

If we have more than one set of operations, we have more than one concept, and strictly there should be a separate name to correspond to each different set of operations. (1927, p. 10)

For Bridgman, it is not the case that there is some construct ‘intelligence’ that multiple intelligence tests measure, and therefore define. Each different intelligence test measures a slightly different intelligence construct.

It should be specified, however, that Bridgman did not always advocate for being this stringent about distinguishing between different constructs in practice. He allowed for the practice of using the same term across multiple measurement methods/concepts, but only in cases in which the various measurement methods were highly consistent (1927, p. 16). However, Bridgman concession was justified only by practicality, and he called for caution when permitting such convergence between concepts. He worried that using the same term as shorthand for multiple different concepts across multiple measurement methods would allow scientists to slip into thinking that a concept could have meaning beyond its concrete measurement operations. Bridgman cautions, “our verbal machinery has no built-in cutoff” (1959, p. 75).

Operationalism is often compared to Logical Positivism due to the similarities between the verifiability criterion of meaning and Bridgman’s operational criterion. According to the verifiability criterion of meaning, a proposition is cognitively meaningful if and only if
it can be proven or falsified by direct experience. Thus both theories grounded meaning in observation. Some logical positivists were early proponents of Bridgman’s work. Most notably, Herbert Feigl (1902-1988), a member of the Vienna Circle, introduced the view to psychologists working at Harvard in the 1930’s (C. Green, 1991). However as philosophers like Carnap and Hempel developed damning criticisms to the verifiability criterion of meaning, their criticisms would also be applied to Operationalism.

3.2.1 Strict Operationalism as a Theory of Construct Validation

Construct validity is simple for strict operationalists. Once a construct is defined by establishing its measurement procedures, the measure has construct validity. There is no coordination problem between measure and construct because measure and construct are the same. The question “What is $x$?” and “How do we measure $x$?” have the same answer and are answered by the same process of operationalization. Moreover, on this view constructs are grounded in reality as they are defined purely by the concrete operations that bring about directly observable phenomena.

The straightforward nature of construct validation may initially seem like a benefit of the view. However, many useful theoretical constructs cannot be operationally defined because of ambiguities about the notion of operation and the restrictive nature of the operational criterion of meaningfulness (Chang, 2009b) For example, Hempel (1954) argued against operationalism because it was unable to define dispositional terms (e.g., ‘solubility in water’). Dispositional terms do not designate currently observable properties, but rather tendencies towards certain behavior under specific conditions. Solubility in water cannot be synonymous with its operations, because the term denotes a potential, rather than an actual, state of affairs:
A concept thus characterized is not “synonymous with the corresponding set of operations.” It constitutes not a manifest but a possible character, namely a disposition to exhibit a particular characteristic response under specified test conditions. (p. 126)

Defining solubility in water as a set of possible observations means that the definition is true in virtue of general laws of nature, not true in virtue of empirical observation (p.127). The only way for operationalism to permit dispositional theoretical terms is to abandon or modify the verifiability criterion of meaning.

Many psychological constructs, especially personality traits, behave more like dispositions than currently observable properties. Thus if we cannot operationally define dispositions, Operationalism is not a tenable position for psychological construct validation. Take narcissism, for example. An Operationalist definition of narcissism would define it as a set of operations, like scoring above a particular threshold on a personality inventory. When ascribed to a person, narcissism then denotes the property of scoring a particular way on a particular test. However, this misses the entire purpose of ascribing personality traits to individuals. When we describe an individual as a narcissist, we are assigning them the propensity to behave in a particular fashion across different contexts, not merely the propensity to score high on a narcissism test. Not thinking of personality traits as dispositions prevent them from fulfilling their primary role: denoting potential, not actual, behavior.

One can respond to this criticism that it requires an uncharitable reading of ?operations?, and that Bridgman himself never denied operational definitions that included potential states of affairs. All operational definitions involve potential states of affairs by designating a set of potential operations; thus dispositions are no different. A cursory search of the psychological testing literature may make it seem that some psychologists believe that dispositions can be operationally defined (see, for example, the NCATE 2000 Standards that purports to
This rebuttal is untenable because it requires misunderstanding dispositions. If narcissism is a disposition, it can be ascribed to an individual even if they are never put in a situation in which they behave in accordance with this personality trait. In other words, narcissism is a disposition because it does not have to ever obtain in order to be accurately ascribed to an individual. However, non-dispositional properties cannot be ascribed if they do not obtain, i.e., are true of the current state of affairs. It makes little sense to claim a 3 ft tall child is, in fact, 6 ft tall because that is their predicted height once they reach adulthood. Therefore, operationalized constructs are not dispositional merely in virtue of being operationally defined.

While misplaced, this rebuttal reveals an additional problem with Bridgman’s view. His notion of operations is ambiguous at best. The common intuition of scientists interpreting Bridgman’s view is that “the operations that matter are measurement operations involving physical instruments” (Chang, 2009b). For example, Tolman, a noted advocate for operationalism in psychology, proposed an operationalization of hunger that defined it as “time since last feeding” (Feest, 2005, p. 136). However, Bridgman (1938) did not intend to restrict operational definitions to laboratory settings. He later wrote that this was the “most widespread misconception with regard to the operation technique” (1938, pp.122-4). In his later work, he distinguished between three types of operations, of which instruments were only one. The other two, mental/verbal, and paper and pencil, were not necessarily laboratory operations (Bridgman, 1959, p. 3). He maintained even in later work that instrumental operations were always preferable, even though he did not justify this epistemic claim. Indeed, Bridgman admitted later that he had not given “an analysis of what it is that makes an operation suitable”, or “in what terms can operations be specified” [p. 77].
Putting aside the difficulties in understanding operations, the core issue with operationalism is in the claim that constructs directly entail a single measurement method. As Hempel points out that if a single set of operations defines each construct, the conceptual unity of theoretical constructs is impossible (1954, p. 91-7). Each different measurement procedure for intelligence, in fact, measures a difference intelligence construct. There is no one “intelligence” construct, but rather a different intelligence construct for each measure: a Stanford-Binet intelligence, a Raven’s Progressive Matrices intelligence, a Wechsler Adult Intelligence Scale intelligence, and so on. Bridgman’s operation criterion leads to “a proliferation of concepts of length, of temperature, and of all other scientific concepts that would not only be practically unmanageable, but theoretically endless” (1954, p. 97). The operational criterion amounts to a view of test validity in which comparative claims between measures are in principle impossible.

Thus strict operationalism is ill-suited for psychology despite its straightforward solution to the coordination problem. The very thing that solves the coordination problem (positing a direct entailment relationship between construct and measurement method) impedes strict operationalism from handling psychological constructs, which are often dispositional (like personality traits) and measured in multiple ways (like intelligence).

3.3 Psychological Operationalists

Although Bridgman was himself a physicist, his ideas were most popular among early research psychologists. These psychologists found operationalism attractive because it prescribed a method of defining hard to define psychological constructs that were not directly observable and steeped with ambiguities from their folk notions. The first explicit formulations of psychological operationalism are by Stanley Stevens (1936) and Edward Tolman (1936).
Stevens was introduced to operationalism while writing his dissertation under Edwin Boring (1886-1968) green1991. Tolman most likely was first exposed to operationalism by Moritz Schlick. Shlick, a founder of the Vienna Circle, invited Tolman to Vienna while he was lecturing at Berkeley as a visiting professor (ibid.). B.F. Skinner, also a student of Boring, never called himself an operationalist, preferring the term “behaviorist” instead. However, his behaviorist views were influenced by strict operationalism. For Skinner, behaviorism was “nothing more than a thoroughgoing operational analysis of traditional mentalistic concepts” (1945, p. 272).

How psychologists interpret operationalism varies widely. This variation is most evident by the stark disagreement between psychologists attending the symposium on operationalism, organized by the Psychological Review (Boring, 1945). The following chart indicates the different responses given to a set of question about Operationalism:
<table>
<thead>
<tr>
<th>Questions</th>
<th>Boring</th>
<th>Bridgman</th>
<th>Feigle</th>
<th>Pratt</th>
<th>Skinner</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purpose of operational definitions</strong></td>
<td>Precision and unification of science</td>
<td>Precision</td>
<td>Eliminate obscurity, ambiguity, vagueness; adapt term to wider context; add new term</td>
<td>Heuristic value</td>
<td>Precision</td>
</tr>
<tr>
<td><strong>Do they form an infinite regress?</strong></td>
<td>Regress ends when agreement is reached</td>
<td>Note infinite; semi-convergent</td>
<td>Not infinite; ends when definition is ‘ostensive’</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Can different operations define the same construct?</strong></td>
<td>Only when the equivalence of the operations is established</td>
<td>Never safe to assume equivalence</td>
<td>Established empirical laws give rules of equivalence</td>
<td>–</td>
<td>Only if the operations result in the same response</td>
</tr>
<tr>
<td><strong>Can operations which are now impracticable be useful?</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Are operations which define non-existent constructs useful?</strong></td>
<td>Yes</td>
<td>?</td>
<td>?</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Can operations which are impossible in principle be useful?</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Question</td>
<td>Yes</td>
<td>Answerable by experiment only</td>
<td>No</td>
<td>Answerable by experiment only</td>
<td>No</td>
</tr>
<tr>
<td>--------------------------------------------------------------------------</td>
<td>-----</td>
<td>--------------------------------</td>
<td>----</td>
<td>------------------------------</td>
<td>----</td>
</tr>
<tr>
<td>Is ‘experience’ a proper construct for operational definition?</td>
<td>Yes</td>
<td>‘Useful’ and ‘non-useful’ are preferable terms</td>
<td>Yes(?)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Are there ‘good’ and ‘bad’ operations?</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Is operationism more than a renewed emphasis on experimental method?</td>
<td>Yes(?)</td>
<td>No</td>
<td>No</td>
<td>All theory is tautological</td>
<td>–</td>
</tr>
<tr>
<td>Does operationism relegate all theorizing to metaphysics?</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>If the operation defines a construct, how can we speak of ‘improving’ an operation (e.g. IQ test)?</td>
<td>By establishment of equivalence to something else</td>
<td>Only the start of a process of spiral approximation</td>
<td>Only the start of a ‘long labor of adjustment and re-definition’</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Must all scientifically legitimate definitions be operational?</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>What is a definition?</td>
<td>A statement of equivalence</td>
<td>Always operational</td>
<td>A rule concerning the use of a symbol</td>
<td>–</td>
<td>the science of verbal behavior of scientists is not yet developed</td>
</tr>
<tr>
<td>Can a concept be defined in terms of the operations which produce it or result from some hypothetical phenomenon?</td>
<td>If correlation between the antecedent and consequent operation is found</td>
<td>Perhaps, but only useful in initial stages, if at all</td>
<td>Yes, initially, but the concept should be well-enough developed that at least two operational definitions are known</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Adapted from Boring (1945)
As evidenced by the above chart, what psychologists took operational analysis to involve significantly varied. Moreover, none of the psychologists writing about operationalism at the time adopted Bridgman’s strict operationalism formulation. For my present purpose, interpreting operationalism as a possible epistemic view about the validation of measures requires an analysis of the different varieties of operationalism psychologists themselves employed.

There are two forms of operationalism I will examine here: *methodological operationalism* and *convergent operationalism*. Feest (2005) attributes methodological operationalism to the first research psychologists writing about operationalism. The view takes a pragmatic approach to operationalism, interpreting it as a method for defining psychological constructs in order to enable empirical investigation rather than a theory of meaning. Convergent operationalism originated among psychologists in the 1950s. Influenced by the critiques of logical positivism’s verifiability criterion of meaning, convergent operationalists believed that a construct’s meaning was synonymous with the convergence of multiple measurement methods. This view intended to avoid the proliferation of concepts, while maintaining operationalism’s reliance on observation.

### 3.4 Methodological Operationalism

I will focus first on Steven’s and Boring’s formulations of Operationalism. Feest (2005) argues that Steven, along with his contemporary Tolman, were methodological Operationalists. Rather than taking Bridgman’s view literally, such that a concept is defined only by its operations and nothing else, methodological operationalists took the operational analysis of psychological concepts to be a scientific tool that enabled empirical research. Evidence for this view is Steven’s claim that psychological constructs have empirical meaning only if they
stand for definite, concrete operations (S. S. Stevens, 1935, p. 517). Thus methodological Operationalists took operational analysis as a necessary criteria for empirical investigation, not for existence itself. Feest writes, “scientists were partially and temporarily specifying their usage of certain concepts by saying which kinds of empirical indicators they took to be indicative of the referents of the concepts.” (p. 133).

As a methodological Operationalist, Stevens primary departure from Bridgman is in taking operational definitions to be dynamic, e.g. subject to revision, and non-exhaustive:

[Definitions] take into account the state of factual knowledge at a given time. It is for good reason that the discovery of new related facts may make a revision of the criteria necessary so that we may include or exclude the new observation from the class denoted by the original definition. . . . No concept can be defined once and for all: every concept requires constant purging to keep it operationally healthy. (1935, pp. 519, 527)

He departs even further by proposing that the fundamental operation was discrimination (1939). Here, he directly contradicts Bridgman, who denied any hierarchy of operations in which one type of operation would be more fundamental than another. Stevens argument was that operational analysis of a concept requires going back until we reach operations which are “unanalyzable and apprehendable only intuitively by personal experience.” (S. S. Stevens, 1939, p. 521). He took experience to ultimately be discrimination, as discrimination is the reportable response to stimuli that is verifiable and repeatable. Discrimination, then, is the only objective part of experience and thus the only part of experience that can be empirically investigated (pp. 521-2). Therefore, discrimination is the fundamental operation.

Edwin Boring (1886-1968) is now infamous for his often quoted statement “intelligence is what intelligence tests test” (1923). This quote is presented out of context often as an example of the inherent absurdity of strict operationalism about psychological constructs.
Yet, Boring’s paper is an informative presentation of methodological operationalism. For our present purposes, it provides a useful example of an account of methodological operationalism that is not restricted to discrimination-based tasks.

Boring’s experience with intelligence testing started during the First World War. Boring led a team of psychological examiners at Camp Upton on Long Island. They developed and administered intelligence tests to military recruits. In 1918, Boring and his team released a massive report on the Army’s intelligence program. This report influenced intelligence testing for decades after. After his experience administering intelligence tests, Boring was concerned with the lack of objectivity in intelligence testing.

In “Intelligence as the Tests Test It”, Boring seeks to explain the psychological construct “intelligence”, not the more broad folk notion. His account includes three central claims: 1. Intelligence is what intelligence tests test, 2. This definition can be revised based on criterion analysis, and 3. There is no direct measure of generalized intelligence. His full definition of intelligence is:

. . .intelligence as a measurable capacity must at the start be defined as the capacity to do well in an intelligence test. Intelligence is what the tests test. . .Measurable intelligence is simply what the tests of intelligence test until further scientific observation allow us to extend the definition.(Boring, 1923)

There are a few important things to note about this definition. By defining the psychological construct “intelligence”, and distinguishing it from the folk notion, Boring rejects Strict

---

4It should be noted that Boring was a highly prolific experimental psychologist outside of his early work on intelligence testing. Besides his early work and writing on intelligence testing, he also worked on misperceptions in visual perception. At Harvard, Boring was the dissertation advisor of Stevens and is credited with first introducing Stevens to Operationalism. Some even credited his 1923 paper as being the first formulation of psychological Operationalism (C. D. Green, 1992).
Operationalism which would claim that intelligence, not just the psychological construct, is operationally defined. Moreover, Boring believes that further scientific observation may permit revisions of his definition. I take this as an indication that Boring endorsed a type of methodological operationalism, in which operational definitions are necessary for initial empirical investigation but are subject to revision.

For Boring, extending the definition of intelligence involves connecting new criteria or capacities to intelligence scores by producing a high enough correlation. If scores on an intelligence test highly correlate with, for example, academic performance, we can then extend the definition of intelligence to include the claim that academic performance is partly dependent upon intelligence. Partly dependent, because less than perfect correlation means that intelligence does not wholly determine academic performance.

3.4.1 Methodological Operationalism as a Theory of Construct Validation

Looking at Steven’s methodological Operationalism, what he takes to be operational analysis is essentially his method of construct validation. Steven’s 1935 paper, “The Operational Definition of Psychological Concepts” provides an informative explanation of operational analysis. In this work, Steven explains that operationally defining a particular concept requires determining which discriminatory act responds explicitly to the concept in question. Doing so is an empirical process that requires experimental tests to determine what discrimination responds to which attributes:

The specification of the attributes of sensation in a given sensory modality and the establishment of adequate criteria for them are contingent upon experimental discoveries. Thus, in the case of audition, the classical criteria of inseparability and
independent variability must be modified not that the manner of the dependence of the attributes upon the dimensions of the stimulus has been discovered, for the relationship turns out to be such that no single attribute can be varied independently while all others remain constant. However, it is possible to hold one attribute constant and vary all the others. By this procedure four aspects of tone, pitch, loudness, volume, and density, have been determined as satisfying the criteria of a tonal attribute. (p. 525)

What Stevens is describing here is a process of isolation, in which one systematically varies different aspects of a stimulus (in this case an auditory tone) and records a subject’s discriminatory response. For an aspect of the stimuli to be an attribute, it must be capable of being held constant while other attributes of the stimuli are varied, and subjects can discriminate between the variations of the attribute in isolation from the others. Stevens describes this as a process of tuning the subject to the different aspects of the stimulus:

The class of reaction which we call attributes is obtained from organisms which have been “set” or “tuned” so that they respond to a certain aspect of the stimulus-process. This tuning of the observer is one of the fundamental operations underlying the concept of attribute. The ability of the experimenter to set the observer, for example, to respond to loudness and not to pitch is crucial to the determination of the attribute loudness. (p. 525)

It is notable that in Steven’s explanation, the very act of discrimination is not explicitly defined. Here Stevens has explained how one isolates an attribute of a stimulus via the discrimination acts of an observer. This explanation leaves open the question as to what makes discrimination a valid measure of some attribute. We can infer from Steven’s requirements for objectively that a necessary criterion for discrimination is that it is verifiable through comparison to other subjects. However, that does not say anything about what would make one form of discrimination a better measure compared to another, or if such comparison is even possible.
Fortunately, Stevens wrote extensively on measurement scales. In “On the Theory of Scales of Measurement,” Stevens introduced a system of scales that differentiated between four types (1946). These four types—nominal, ordinal, interval, and ratio—are still used today. In the paper, Steven defines measurement as “the assignment of numeral to things so as to represent facts and conventions about them” (p. 680). He goes on to write:

Measurement is never better than the empirical operations by which it is carried out, and operations range from bad to good. Any particular scale, sensory or physical, may be objected to on the grounds of bias, low precision, restricted generality, and other factors, but the objectors should remember that these are relative and practical matters and that no scale used by mortals is perfectly free of their taint. (p. 680)

In other words, for something to be a measurement method requires that it assigns numerical values to things in a manner that represents facts and conventions about them. A measurement method is better or worse depending on its potential bias, precision, or scope. There is the additional criterion that the measurement method utilizes a type of scale that accurately represents features of the measured quantity. Stevens states these criteria as a question of the mathematical group-structure of the scale form: “In what ways can we transform its values and still have it serve all the functions previously fulfilled” (p. 680). Thus, it would be inappropriate to assign an ordinal scale to color categories, such that red = 1, blue = 2, green = 3, and so on, because the mathematical operations possible with such a scale do not accurately reflect facts about the relationships between the different colors. Red + blue becomes 1+2 = 3; however, it is not the case that combining red and blue creates green.

For methodological operationalists, construct validation consists of first operationally analyzing the concept to be measured, and then the assignment of a scale to the operation(s) that empirically define the concept. Later, the revision of a measurement method happens when
the type of scale used is changed to reflect better the relationships between the measured quantity.

Boring provides an additional step of potential construct revision: the extension of a definition though correlational analysis with observable criteria. Note that this is a revision of a construct’s definition, not the measurement method itself. However, it seems plausible that in elaborating various relationships between a construct and criterion, researchers develop new ways of measuring the construct. If intelligence, for example, correlates highly with academic performance, then presumably measures of academic performance can serve as indirect measures of intelligence (and vice versa).

Taking Stevens and Boring together, construct validation for methodological operationalists goes as follows:

1. Initial operational analysis of the construct that results in an operation/set of operations

2. Assignment of a scale to the operation(s)

3. Extension of the definition through correlational analysis with observable criterion

While methodological operationalism offers a much more developed account of construct validation than strict operationalism, it is incomplete/inadequate as a method of construct validation beyond the initial creation of a measurement method. What is missing is any prescription of how to evaluate and revise the content of a measurement method itself. Definition extension, as described by Boring, merely draws connections between the construct and other observable criteria. This process amounts to descriptive science. The methodological operationalists could, starting with their initial operational definition of intelligence, conclude that occupational success is in part determined by intelligence if intelligence test scores
highly correlate with occupational success. However, methodological operationalism does not account for revisions of measures based on the failure to find predicted correlations, i.e., hypothesis testing. This failure is apparent when one tries to apply methodological operationalism to early revisions of the Stanford-Binet Scale of Intelligence.

David Wechsler (1896-1981) realized during his time as a testing psychologist in the U.S. Army that many of the men he tested using the Stanford-Binet Scale were scoring as low-functioning even though they had no difficulties functioning as successful adults before being drafted into the Army (Boake, 2002). Thus there was a discrepancy between the general capabilities demonstrated by some men in their every-day lives, and their predicted functioning based on the scores they received on the intelligence test. In order to resolve this discordance, Wechsler developed a new test, the Wechsler-Bellvue Scale, that while used similar test items as the Stanford-Binet, heavily revised its original scoring procedures (1939). His motivation was not to create a new test that tested a different type of intelligence than the Stanford-Binet, but instead, he intended to create a measure that better captured the notion of intelligence that both tests meant to measure. This motivation is evident when he justifies his use of test items from already existing intelligence tests (rather than creating new items from scratch):

. . .[my] aim was not to produce a set of brand new tests but to select, from whatever source available, such a combination of them as would best meet the requirements of an effective adult scale. (p. 78)

Wechsler’s revision changed the content of the test, removing some test items while adding others (Boake, 2002). He also revised the scoring procedures for the test. However, he did not change the type of scale used. While the new test reported a point scale score rather than a mental age, both types of scoring systems used an interval scale. The primary difference in scoring was that Wechsler’s test omitted the categorization of test takers into “mental age
groups” based on their final scores. Wechsler’s contribution, then, was to change the content of the test in order to create a test that correlated with the criterion he believed intelligence should relate to.

Methodological operationalism permits test revision, but only explicitly accounts for two types: scale revision or definition extension. Scale revision involves changing the type of scale used, in order to more accurately represent the relations between measured quantities, or revision of the scale in order to represent changes in a single quantity better. Wechsler’s revision appears to be of the second type. While he did not change the type of scale used, he revised the scale so that it would more accurately represent what he saw to be the real variation between individuals’ intelligence. Wechsler’s scale revision then seems to fit methodological operationalism’s account.

Wechsler’s content revision, on the other hand, cannot be accounted for by methodological operationalism. Definitional extension, according to Boring, is a practice of observation. When one observes a high enough correlation between test scores and some criterion, then the definition of the construct the test measures can be extended to the new criterion. It may seem like what Wechsler did was extend intelligence to a new criterion. However, he was engaged in hypothesis testing, not observation. He initially failed to find the correlations he expected between intelligence scores and individuals’ success in life. It was this failure to confirm a hypothesized relationship between intelligence and observable criteria that drove his revision of the test. Methodological operationalism cannot account for this type of test revision, and therefore is an incomplete account of construct validation.
3.5 Convergent Operationalism

Convergent Operationalism is the view that the meaning of a construct is determined by the convergence of multiple measurement methods (Garner, 1954; D. T. Campbell & Fiske, 1959), the psychologists who developed the first statistical measure of construct validity and distinguished between the two types of adequacy discussed in the previous chapter, were convergent operationalists. Campbell, in particular, advocated for this more permissive version of Operationalism throughout his work, though what he named his view varied over time. The names he used included methodological triangulation (1959), operational delineation, (1954), and convergent validation (1957).

Whereas Bridgman’s Operationalism required complete congruence between measurement methods for term confluence, Convergent Operationalism requires only partial convergence. Partial convergence means that the disparate measures are highly correlated, and the correlations are not due to method variance. Method variance is any variance in test scores due to the measurement method itself and not variance in the actual trait/attribute being measured.

Convergent Operationalism is attractive to psychologists because it permits unobservable psychological constructs: “The importance of the method of converging operations is that it provides the only means of making potentially valid inferences about unobservable constructs or processes” (Grace, 2001, p. 28). This is important for psychologists who reject strict behaviorism and want to permit mental entities into their psychological theories.

The statistical motivation for this form of operationalism is Classical Test Theory, which was widely assumed by statisticians in the time Campbell and Fiske wrote. According to Classical Test Theory, the application of a measurement method produces an observed score comprised of two parts, the true score and error: observed score = true score + error (Novick, 1966).
Method variance is one potential source of error. If there is only one method of measuring a construct, there is no way to estimate how much of the observed score is due to method variance. Thus, for statisticians like Campbell and Fiske, who believed every observed score had some component of error, strict operationalism was an untenable view because it denied the possibility of such error. Convergent operationalism offered an alternative whereby they could maintain a strict direct entailment relationship between measure(s) and construct, while endorsing Classical Test Theory.

Convergent operationalism at first seems to fair better than methodological operationalism. Convergent operationalism enables comparison between measures of the same construct, thus satisfying the first criterion and avoiding Hempel’s proliferation of concepts worry. Measures can be compared based on a variety of statistical measures, including Campbell and Fiske’s (1959) multi-trait multi-method matrix. For example, we can determine that one intelligence test is better (i.e. more construct valid) than another if the test is more reliable, or is less influenced by method variance.

However, further analysis of how convergent operationalism grounds the meaning of theoretical concepts reveals that this view leads to a dilemma. Either it is the case that convergent operationalists believe that constructs have meaning beyond their measures, and are not truly operationalists at all, or they do not have a satisfying account of what makes convergence between measures meaningful.

Convergent operationalism takes the operational criterion of meaning, in which a construct is “synonymous with its set of operations”, and modifies it. The modified criterion is a convergence criterion of meaning, in which a construct is “synonymous with the convergence of multiple measurement methods”. Thus, while Bridgman would claim a construct is
nothing more than its corresponding set of operations, convergent operationalists claim that a construct is nothing more than what the measurement methods converge on.

This notion of ‘convergence’ should be treated with suspicion. To define intelligence as “the thing that all intelligence tests converge on” invokes some object, ‘intelligence’, that exists apart from the test. “Intelligence” is then understood as a concept that underlies the various intelligence tests. This definition, of course, goes in direct contradiction of the motivation for operationalism, which is to eliminate talk of concepts above and beyond direct observations. Alternatively, if the convergent operationalists maintains that what is converged upon by multiple tests does not exist apart from its measures, they are then faced with the problem of explaining which convergence is meaningful and which is not. It is unlikely that all convergence between measures is meaningful. As Green (1992) points out, convergent operationalism creates the real possibility of vacuous convergence:

If there is no concept “apart from” the operations used to measure its presence, then there is, to put it in its bluntest form, nothing for the operations to converge upon. Mightn’t the operations merely converge upon each other? Of course, but without an underlying concept relating them, this would have no more significance than both a thermometer in the mouth and a tape measure around the belly both reading ‘37’.

In other words, unless one allows for concepts to exist apart from their operations, convergent operationalism cannot determine what measure convergence is meaningful (i.e., reflects a shared construct) and is accidental. But if one rejects that constructs are nothing more than their operations (be it a single measurement method, or many), one is no longer an operationalist of any form.

Moreover, by adopting convergence as the ultimate measurement goal, convergent operationalism faces the same discordant data problem Stegenga (2009) voiced about robustness.
Convergent operationalism leaves it as an open question whether discordant measures must be revised in order to maximize concordance, or if theories may be revised in order to maximize concordance. Without some account of how one reasons about this choice, it requires stipulation either about the correctness of one’s measures or the correctness of one’s theory.

### 3.6 Conclusion

Operationalism played a critical theoretical role in early research psychology, guiding initial efforts by psychologists to measure non-observable mental processes and entities. However, as a theory of construct validation operationalism fails, even if one adopts one of the various reinterpretations of the view. The various issues I have examined here may be summed up as follows: operationalism incorrectly claims that the relationship between construct and measure is direct entailment. Bickhard (2001) echoes this sentiment, writing:

> Measuring is important; being precise about measurement procedures is important; reliability and validity are important; testing hypotheses and theories is important; scientific meaning is important; theories are important. But none of them can be equated to any of the others. Yet operationalism conflates all of them, thus muddying our thinking about these issues. That is the tragedy of operationalism. (p. 8-9)

If measure and construct directly entail one another, then one answers the question “What is $x$” by answering the question “How do we measure $x$”. But, as I argued in the previous chapter, construct validity and construct legitimacy, while related, respond to different types of evidence. In the end, operationalism in all its forms fails to respect this distinction. By conflating both questions operationalism is unable to account for the vast majority of
psychological research in which a single measurement method does not exhaustively capture a construct.
Chapter 4

Hypothesis-Testing

4.1 Introduction

In the previous chapter, I analyzed three versions of Operationalism as potential epistemic theories of justification for construct validation. In this chapter, I turn to the second contender for construct validation: hypothesis-testing.  

Hypothesis-testing refers to the theory of construct validation first advocated for by Cronbach and Meehl (1955). As was the case with Operationalism, the psychologists who first formulated hypothesis-testing drew inspiration from their philosophical contemporaries. Drawing on the work of logical empiricists, in particular, Carl Hempel, Cronbach and Meehl proposed construct validation as a method of validating measures that measured constructs that were not exhaustively defined by their operationalization. Cronbach and Meehl adopted logical empiricism’s syntactical criterion of meaning so that the meaning of a construct was determined by the construct’s place in a broader theory and thus not dependent solely on its operationalization. In adopting

\footnote{A note on my terminology here: “hypothesis testing” here refers to the general practice of testing predictions entailed by a theory and/or model; “hypothesis-testing” refers to a particular theory of construct validation.}
the syntactical criterion of meaning, Cronbach and Meehl also endorsed a version of the syntactic view of scientific theories, drawing from Hempel’s description of a nomological network (Hempel, 1952).

Cronbach and Meehl’s method of construct validation through hypothesis-testing was an attempt at translating Hempel’s framework to psychological measurement. Adopting a syntactic view of scientific theories meant that their method was not merely a general hypothesis testing approach. Rather, Cronbach and Meehl committed to a to a specific, deductivist syntactic view of hypothesis testing. In doing so they failed to account for the myriad of ways in which psychological theories and explanations are discordant with the syntactic view of scientific theories. The syntactic view has been heavily criticized by philosophers going back to the 1940s (Suppe, 1977), so it is unsurprising that hypothesis-testing is inadequate as a theory of construct validation. Yet the spirit of hypothesis testing—indeed independent from any commitment to logical empiricism—is still prevalent among psychologists writing about construct validation (see Smith, 2005).

Borsboom and Mellenbergh (Borsboom et al., 2004) argue that construct validation is a process of empirically testing a hypothesis about a construct (p. 1066). Similarly, Alexandrova and Haybron (2016) describe the implicit logic of construct validation as hypothesis testing. If there is any consensus among psychologists about what construct validation requires, it is that is involves hypothesis testing:

To determine whether a measure is useful, one must conduct empirical tests that examine whether the measure behaves as would be expected given the theory of the underlying construct (Diener, Lucas, Helliwell, Helliwell, & Schimmack, 2009, p. 67)
Indeed, there is something right about the idea that one must test a hypothesis about a construct in order to validate a measure of that construct. I will argue here that while hypothesis testing is an undeniably important component of construct validation, the hypothesis-testing method developed by Cronbach and Meehl is inadequate. Not only is hypothesis testing Moreover, hypothesis-testing interpreted as an epistemic theory of construct validation fails to solve the coordination problem. Finally, even when evaluated independently of the syntactic view of scientific theories, hypothesis-testing mischaracterizes the bidirectional inferential relationship between measures and constructs.

In this chapter, I explain Hempel’s logical empiricist framework that inspired Cronbach and Meehl’s method in section one. I then demonstrate in section two how Cronbach and Meehl applied Hempel’s framework to construct validation. In section three, I analyze hypothesis testing as an epistemic theory of justification for construct validation. I draw from the literature criticizing logical empiricism for inadequately or inaccurately characterizing scientific theories. The biggest issue for hypothesis-testing is that its solution to the coordination problem is inadequate. Coordination, for hypothesis-testing, requires stipulations that are not justified. Moreover, hypothesis-testing introduces a new problem of interpreting discordant data. However, even if hypothesis-testing is evaluated independently of a logical empiricist framework of scientific theories, it still fails as a theory of construct validation due to its mischaracterization of the inferential relationship between measures and constructs.

Hypothesis-testing characterizes the relationship between measures and constructs as a unidirectional, top-down relationship in which the theory of a construct determines what measures are appropriate. This account neglects the influence measures themselves can have on a construct’s theory. In section four, I argue that Hypothesis testing cannot account for bottom-up theory revision. I use Spearman’s (1904) factor analysis of intelligence measures
as an example of such revision and demonstrate how his revision of intelligence theory was driven by a mathematical advancement in statistics, not hypothesis-testing.

4.2 Hypothesis-Testing

Hempel not only levied many compelling refutations of Operationalism but also created the framework that Cronbach and Meehl (1955) utilized in their account of construct validation. Hempel was a logical empiricist, and as such endorsed a syntactical criterion of meaning. According to the syntactical criterion, a construct is defined only in part by observational terms. This is in direct contrast with the operational criterion (from operationalism) or the verifiability criterion (from logical positivism), that restrict meaning to only observational terms. For Hempel, the meaning of a construct is determined not by the operations that measure the construct, but instead by the construct’s place in the broader scientific theory of which it is a part (Hempel, 1952). Thus, determining the meaning a construct requires an account of the structure of scientific theories.

Hempel’s account of scientific theories is a syntactic view, due to his emphasis on the logical structure of scientific theories (Wessels, 1976). The motivating belief for the syntactic view, also known as the ‘received view,’ is that scientific theories are linguistic structures, made up of logical, theoretical, and observational vocabulary (Hempel, 1952; Carnap, 1939; Nagel, 1961; Feigl, 1970; Hempel, 1970). A scientific theory, on this view, is a formal system made up of 1. symbols, 2. formation rules, and 3. deduction rules (Halvorson & Tsementzis, 2015, p. 3). This formal system supports deductions to empirical claims, i.e., observations. Thus, while the syntactic view still maintains that there is a logical relationship between theory and observations, as does Operationalism, this relationship is not one of direct entailment, but rather deduction from the entire theory containing a construct.
The following table summarizes the main components of a theory on the syntactic view:

**TABLE 4.1**

Elements of Scientific Theories

<table>
<thead>
<tr>
<th>Sentence Type</th>
<th>Theory</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terms</td>
<td>Tₘ</td>
<td>Cₘ</td>
</tr>
<tr>
<td></td>
<td>Sₘ</td>
<td>Oₘ</td>
</tr>
<tr>
<td>Language</td>
<td>Lₜ</td>
<td>Lₜ and Lₒ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lₒ</td>
</tr>
</tbody>
</table>

Adapted from Winther, Rasmus Gronfeldt, “The Structure of Scientific Theories”, The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.). Adapted with permission

As scientific theories are primarily linguistic, the main components of a theory for the syntactic View are languages, sentences, and terms. Starting with languages, Hempel (1952) takes there to be two: theoretical language (Lₜ), comprising of theoretical and logical terms, and observational language, comprising of observational and logical terms (Lₒ). There are three types of terms, e.g. vocabulary: logical, theoretical, and observational. Logical vocabulary includes operators from first-order predicate logic (e.g. connectives like ∧, ∨, or →), and quantifiers (e.g. ∃, or ∀).

Extralogical vocabulary is either theoretical or observational. Observational vocabulary is predicates that refer to directly observable entities or attributes. Theoretical vocabulary involves theoretical terms, i.e., constructs. These terms are predicates that refer to entities or attributes that cannot be observed directly, e.g. “mass” or “length”. As theoretical vocabulary refers to non-observables, theoretical terms are not directly defined in terms of observation like observational vocabulary. A theoretical term is instead defined indirectly via correspondence sentences (i.e., bridge principles) that connect theoretical terms to observational terms.

The last component is sentences, which are comprised of strung together terms. Sentences comprised of logical and theoretical terms, Tₘ, are theoretical sentences. These are the
axioms, theorems, and laws of a theory, what Hempel refers to as internal principles (Hempel, 1966, p. 72). CS are correspondence sentences, or bridge principles, that connect theoretical terms to the relevant observational terms. Finally, OS are observational sentences made up of observational and logical terms.

The general idea is that scientific theories are a network of interlocking laws, axioms, and theorems; what Cronbach and Meehl referred to as a nomological network (1955). This network is linked to empirical observation via correspondence sentences that serve as predictive bridges. The key feature of the nomological network is that a construct is defined not just by its observation, but also by its place in the broader theory. A construct’s place in a theory is determined by its relation to other constructs, as determined by the theory’s laws, and its relationship to empirical observations, as determined by the theory’s bridge principles. The bridge principles enable one to derive observational interpretations of a theory, but the bridge principles do not reduce the theory to observational terms. There is always surplus meaning that remains in the theory (Hempel, 1958, p. 87). Thus a construct that cannot be exhaustively defined operationally can still be meaningful if the construct can be indirectly connected to observational predictions, via theoretical definitions and correspondence sentences:

The whole system floats, as it were, above the plane of observation and is anchored to it by rules of interpretation. These might be viewed as strings which are not part of the network but link certain points of the latter with specific places in the plane of observation. By virtue of those interpretive connections, the network can function as a scientific theory: From certain observational data, we may ascend, via an interpretive string, to some point in the theoretical network, thence proceed, via definitions and hypotheses, to other points, from which another interpretive string permits a descent to the plane of observation. (Hempel, 1952, p. 36)
In contrast to operationalism, on Hempel’s view “intelligence” is not defined as “whatever intelligence tests measure”. Rather, intelligence is what intelligence tests measure (if these tests are valid) in addition to whatever the theoretical generalizations that make claims about the connection between intelligence and other constructs (working memory, learning, etc.) and possible observations refer to.

For example, Spearman’s two-factor theory of intelligence claims that intelligence is composed of two factors: the g factor, general intelligence, and the s factor, specific intellectual abilities spearman1927. Spearman’s research on the g factor lead him to make a variety of generalizations about the connection between g and different cognitive abilities. According to Spearman, logic, spatial abilities, language, and mathematical ability are all related such that a high g factor is positively correlated with these abilities. Under the syntactic view, Spearman’s definition of the g factor would include such generalizations that connect intelligence with other cognitive capacities like spatial ability.

The utility of the syntactic view of scientific theories for psychologists is self-evident. While strict operationalism leads to the problematic conclusion that most psychological constructs are meaningless, Hempel’s account permits non-observable constructs. In fact, Hempel embraces non-observable constructs (what he calls “theoretical constructs”) as a necessary component of scientific explanation (1952, p. 31). For a construct to be scientifically admissible, it needs only to occur in the nomological net and that at least some of the laws in the network involve observables. For Cronbach and Meehl, it also points directly to a method of construct validation.
4.2.1 Construct Validation through Hypothesis-Testing

Adopting Hempel’s syntactic view, Cronbach and Meehl formulated a method of construct validation based on hypothesis testing. Their view is that one validates some measure, \( m \), of a construct, \( C \), by using \( m \) to empirically test the hypotheses derived by the theory that contains \( C \). Each successful test using \( m \) adds evidence to the belief that \( m \) has construct validity relative to \( C \). A failure to confirm a hypothesis is evidence against the construct validity of the measure and indicates that either the theory or measure needs revision. What particular types of empirical tests are required to validate a measure depends on the predictions entailed by the construct’s theory.

When psychologists give accounts of construct validation, they still follow the hypothesis-testing framework introduced by Cronbach and Meehl. Haybron and Alexandrova (2016), drawing from contemporary accounts of construct validation (DeVet et al 2011, Simms 2008) describe construct validation as involving three stages. In the first conceptual stage, researchers define the construct to be measured. This definition may not amount to a full theory or model of the construct, however it at minimum needs to describe the scope and limits of the construct. If, for example, the construct is a cognitive capacity, the definition would set what observable behaviors will involve the capacity. In the second stage, researchers chose a measurement method and create the test items, tasks, or other procedures that make up the measure. In the third stage, the measure is tested for its validity.

Haybron and Alexandrova further elucidate what they take to be the implicit logic of tests for construct validation:

A measure \( M \) of a construct \( C \) is validated to the extent that \( M \) behaves in a way that respects three sources of evidence:
1. M is inspired by a plausible theory of C specified in stage 1.
2. Subjects reveal M to track C through their questionnaire answering behavior.
3. Other knowledge about C is consistent with variations in values of M across contexts.

Their description is aimed at the construct validation of self-report measures that consist of multiple test items (i.e. questions) and scoring procedures for the item responses. However, this characterization of the implicit logic can be generalized to any psychological measure by replacing “questionnaire answering behavior” with any type of measurable response, like reaction time, eye gaze location, etc.

What Haybron and Alexandrova highlight here is that the initial development of a measure is driven by a theory of target construct, and is then justified by empirical tests of that theory using the measure. The specific empirical tests are determined by the theory of the target construct. In psychology, the vast majority of construct theories make predictions about consistency of the measured construct in individuals and across constructs.

The previous chapter’s example of Wechsler’s revision to the Stanford-Binet intelligence test exemplifies the hypothesis-testing method, including the function of consistency predictions about target constructs. Wechsler endorsed a theory of intelligence in which intelligence was a general cognitive capacity related to life aptitude. As a general cognitive capacity, intelligence should be consistent across different measures (2) and exhibit similar effects between different populations (3). Consistency across different measures of intelligence is evidenced by higher correlations between scores of an individual on different intelligence tests, and between individual test items. The Stanford-Binet intelligence test met this evidentiary requirement. The problem Weschler observed was in the consistency of variations in intelligence scores across different populations, i.e. contexts.
According to Weschler’s definition of intelligence, intelligence should positively correlate with an individual’s ability to function in society. He predicted a high correlation between scores on the Stanford-Binet and a concurrent criterion: life-success. This effect pattern would be consistent across different populations. However, the data he collected while giving the Stanford-Binet test to army recruits did not support this hypothesis. Time and time again he saw men who functioned perfectly fine in society, who held down jobs and supported families, score low enough on the test to be considered mentally disabled. A particular effect pattern, in this case a high correlation between scores on the Stanford-Binet and life-success, was not consistent across different populations.

Weschler interpreted the failure of the Stanford-Binet test to confirm his prediction as evidence that the test itself lacked construct validity and required revisions. He then developed his test with the aim of producing data that would confirm his hypothesis (Boake, 2002). In doing so, Weschler created a test that met both evidentiary requirements (2 & 3).

4.2.2 Evaluating Hypothesis-testing

Hypothesis-testing has a strong intuitive appeal as a method for construct validation. Unlike the variations of Operationalism previously examined, hypothesis-testing enables evaluative claims about measures. A measure that fails to confirm a construct’s related hypothesis is not construct valid, and a measure that confirms more of a construct’s related hypothesis than another is comparatively more construct valid. Indeed, even the most radical departure from Cronbach and Meehl’s account, Messick, still endorse some basic method of construct validation by hypothesis testing (1987). However, what needs to be evaluated is not merely the method hypothesis-testing, but the epistemic theory of measurement validation upon which it rests. While the general notion of hypothesis-testing may, in fact, be a correct
method of construct validation, the underlying epistemic theory fails to satisfy the criterion relevant to psychological constructs in particular.

There are two related problems here. First, the syntactic view has difficulty accounting for theory revision. As a result, hypothesis-testing permits theory revision, or measure revision, when measures produce data discordant with a theory’s hypotheses. However, hypothesis-testing provides no account of how researchers deliberate between these two options. This problem of interpreting discordant measurement data was discussed in chapter one. Recall that Stegenga (2009) argued that such interpretation required stipulations. Thus there is a real worry here that without some account of how one reasons about discordant data, hypothesis-testing introduces more circular reasoning into construct validation.

Second, hypothesis-testing does not seem to permit theory revision that is not driven by direct tests of a hypothesis. The revision I have in mind is any revision, of measure and/or theory, that is justified not by theoretical considerations, but rather technological and statistical advances that are theory-independent. I refer to this form of theory revision as bottom-up theory revision and will examine this problem in the following section.

4.2.3 Problems with the Syntactic View

Setting aside these two problems with theory revision, it is worth first worth considering the problems with the syntactic view itself, and how these may relate to deductivist hypothesis-testing’s adequacy as a theory of construct validation. Wechsler’s revision demonstrates the four stages of construct validation: 1. Formation of a hypothesis based on a theoretical definition of a construct, 2. Testing the hypothesis through the application of the test one intends to validate, 3. Interpretation of the data as either confirming or disconfirming the hypothesis, and 4. Revision of the test or theory based on the interpretation in step 3.
It may seem like construct validation, explicates as hypothesis-testing, does not require adopting a syntactic view of scientific theories. However, absent of some view of scientific theories, hypothesis-testing has no account of how one forms hypotheses or how one interprets experimental results concerning a hypothesis. In the first stage of hypothesis formation, the syntactic view of scientific theories makes claims about how the theory of a construct is characterized, and further how testable empirical predictions are determined based on the theory. In the final stage, Cronbach and Meehl’s interpretation of the syntactic view leads to commitments about how the failure or success of the empirical tests is to be evaluated.

The first stage requires formulating the theory surrounding the target construct in accordance with the syntactic view, e.g., as a nomological network:

A necessary condition for a construct to be scientifically admissible is that it occurs in a nomological net, at least some of those laws involve observables (Cronbach and Meehl, 1955, p. 290).

While the syntactic view opens the possibility of defining non-observable constructs, it brings with it the high cost of requiring well-developed theories that contain lawful (or for Cronbach and Meehl, law-like) generalizations about the construct.

This is a prohibitively high cost for most, if not all, sciences, let alone psychology. There exist no laws in psychology, or even universal generalizations that could be characterized as law-like (see Rosenberg, 1985, Schaffner, 1993, Smart 1963). Like biology, generalizations in psychology are contingent on how humans have evolved. Thus they are not necessary, nor exceptionless. This was not lost on Cronbach, who later wrote, “it was pretentious to dress up our immature science in the positivist language; and it was self-defeating to say . . . that a construct not a part of a nomological network is not scientifically admissible” (1987, p. 12).
Recall that according to the syntactic view, a scientific theory is an axiomatized deductive system combined with bridge principles that give empirical import to the axiomatic system (Hempel, 1952; 1966). In order for a theory to be axiomatized, it must contain laws or empirical generalizations. Further, on the syntactic view, scientific theories explain in virtue of these laws:

Theories are normally constructed only when prior research in a given field has yielded a body of knowledge that includes empirical generalizations or putative laws concerning the phenomena under study. A theory then aims at providing a deeper understanding by construing those phenomena as manifestations of certain underlying processes governed by laws which account for the uniformities previously studied, and which, as a rule, yield corrections and refinements of the putative laws by means of which those uniformities had been previously characterized (Hempel, 1966, p. 142)

While there are many different philosophical accounts of what exactly a law is, Hempel’s characterization is sufficient for our present purposes. For Hempel, laws are true statements of universal generalization that are logically contingent and unlimited in scope (1966, ch. 5). Note that laws are not just any true universal generalizations, as this permits accidental generalizations. A law is additionally a true generalization that supports counterfactual conditionals (p. 56).

Fodor (1989) argues that in the special sciences, e.g., any science that is not fundamental physics, a generalization will most likely have limiting conditions such that they are not universal, or exception-less. These ceteris-paribus laws are generalizations that hold only with provisions. For example, any generalization made by psychology to predict or explain human behavior is stipulated with an implicit ceteris paribus clause that limits the generalization to humans. Thus, according to Fodor, psychology cannot, in principle, provide generalizations that are unlimited in scope.
One may respond to Fodor’s criticism by weakening the law requirement of the syntactic view. Laws, then, must not be unlimited in scope, but rather can be limited as long as they contain the relevant ceteris paribus clause. Hempel et al. (1965, pp. 166-7) voices a general concern that in stipulating a law’s ceteris paribus conditions one runs the risk of creating a meaningless truism. By qualifying a generalization with too general a ceteris paribus clause, the theory holds unless it does not hold. Then, no empirical evidence could ever disconfirm the generalization as that evidence could be attributed as a violation of the ceteris paribus clause.

Now, one might think this criticism requires an uncharitable reading of Cronbach and Meehl. At other times, they imply that their notion of ‘law’ is not nearly as restrictive as early logical empiricists’. Later on in their paper, they write:

. . . a rigorous (though perhaps probabilistic) chain of inference is required to establish a test as a measure of a construct. To validate a claim that a test measures a construct, a logical net surrounding the concept must exist. (Cronbach and Meehl, 1955, p. 291; emphasis added)

In this passage, they seem to be employing a different revision of the law provision, in which laws include probabilistic generalizations. However, even if we adopt this more permissive application of the nomological network, hypothesis testing still requires standards that most psychological constructs will fail to meet.

There are philosophers that disagree with Hempel, and offer accounts of laws that permit ceteris paribus clauses that directly tackle his triviality worry. Woodward’s invariance theory of laws and Lange’s stability theory of laws both permit ceteris-paribus generalizations as laws, especially in the special sciences (Lange et al., 2000; Lange, 2002; Woodward, 2000b, 2002). These theories take generalizations with ceteris-paribus clauses to be laws when the generalization is stable or invariant over a limited range of contexts. A ceteris-paribus law is then understood as a counterfactual, i.e. an ‘if-then’ assertion, that yields a reliable inference rule. These accounts, however, require abandoning the deductive hypothesis-testing account.
The main feature of psychological constructs I highlighted at the beginning of the chapter is that they are frequently under-defined. If a measure is validated against an under-defined construct, the available hypothesis that can be tested are limited. Hypotheses in these cases are limited in the sense that they are imprecise, or quite literally limited in number. Theories for which probabilistic hypotheses can be derived do exist in psychology, especially in cognitive psychology in which models of cognitive capacities often represent multiple sub-capacities with estimations of the strength of correlation between variables (e.g., structural equation modeling). However, these models exist in well-developed fields, where measures have already been developed and validated. Such models do not exist for new research projects. It seems then that according to the hypothesis-testing approach, construct validity requires putting the cart before the horse—having a well-developed theory of our target construct in order to validate a measure of that construct. Of course, this brings us face to face with the coordination problem. We cannot develop such theories of a construct unless we have valid measures of the construct with which to test our empirical predictions.

4.2.4 Bootstrapping and the Discordant Data Problem

Cronbach and Meehl (1955) were not oblivious to this shortcoming. In fact, part of the motivating worry for staunch Operationalists like Anastasi was that validating measures of non-observable (i.e., not defined exclusively by criterion) constructs would require too much speculation about the nature of the construct. Rather than develop a method that could accommodate these constructs, Anastasi advocated for avoiding them altogether. Cronbach and Meehl’s reply to Anastasi is directly relevant to the under-defined worry. They propose that psychologists “lift themselves up by their bootstraps”. Using the example of the standardization of temperature measurement they write:
This whole process of conceptual enrichment begins with what in retrospect we see as an extremely fallible 'criterion,' the human temperature sense. That original criterion has not been relegated to a peripheral position. *We have lifted ourselves by our bootstraps,* but in a legitimate and fruitful way. (1955, p. 286, emphasis added)

Later on, they add:

When a construct is fairly new, there may be few specifiable associations by which to pin down the concept. As research proceeds, the construct sends out roots in many directions, which attach it to more and more facts or other constructs. Thus the electron has more accepted properties than the neutrino; numerical ability has more than the second space factor. (p. 291)

It seems then that Cronbach and Meehl have in mind a gradual concept-enrichment process, that starts with some instrumentally valuable stipulations about the construct that may later be abandoned as more data is collected. In other words, boostrapping.

Glymour (1980) provides a ‘bootstrapping’ model of confirmation that initially seems helpful here. His account shows how evidence can be applied to a theory in a non-circular way even when the target theory is itself used to establish the evidentiary relationship between the hypothesis and evidence. Boostrapping inferences are non-circular if, “using the theory, we can deduce from the evidence an instance of the hypothesis i.e., an hypothesis comprising or instantiating the test theory, and the deduction is such that it does not guarantee that we would have gotten an instance of the hypothesis regardless of what the evidence might have been.” (p. 127). An important requirement of his account is that the hypotheses specify a determinate value for the relevant variables. Van Fraassen argues (1983, p.32-33) that young sciences, like psychology, rarely provide determinate values, or even ranges of values.
Wylie (1986) provides a more accommodating account of bootstrapping, what she calls ‘constructive bootstrapping’. On her account, bootstrapping is “as much a process of theory construction as it is of theory testing.” (p. 315) Arguments for the evidentiary relevance of specific data requires relying on background knowledge and related theories. These inferences are parasitic in that they require “substantive considerations of content” about relevant variables from other theories. Bootstrapping in psychology then depends on the ability of related fields to provide that content. Whether they do is an open question.

While this is a potentially promising account of non-circular bootstrapping of confirmatory evidence, there remains the question of how to interpret non-confirmatory, e.g. incongruent, data. An account of interpreting incongruent data is essential, as this is the data that is potentially evidence for revision of either one’s theory or measure. Without such, the coordination between measure and construct is still circular. The problem is merely shifted from the onset of construct validation (the stipulation of an initial account of a construct by which we can test a measure) to the interpretation of the data produced by the measure being validated. When the empirical application of a measure produces data that is inconsistent with our held definition of the construct, do we interpret this as evidence our definition needs to be modified, or as evidence that the measure is less construct valid? Either interpretation requires a stipulation. On the one hand, if we interpret the data as evidence that our measure is less construct valid, we are assuming our definition (and its predictions) is correct. On the other hand, if we modify our definition of the construct, we are assuming the measure has some sufficient degree of validity.

This is the worry voiced by Stegenga (2009) in regard to interpreting discordant data across measures. Identifying the source of error, either with the measurement method or the theory, cannot be done without a stipulation. There are two solutions we can adopt in response to this problem. On the one hand, we can completely avoid any stipulation by rejecting a notion
of construct validity that applies to any case in which the construct is not completely defined. On the other hand, we can adopt some type of Conventionalism in which we embrace the necessary stipulations.

Completely defined in this context means that a mechanism that causally links the scores produced by a measure and the construct has been described. This is, essentially, the criteria from Borsboom’s (2004, 2009) theory of construct validation. Recall that on Borsboom’s account that tests like the IAT, which measure undefined (i.e., ‘lumpy’) constructs, cannot have construct validity. If we do not know what implicit attitudes are, to the degree that any failure of a measure’s application can only be interpreted as evidence that the measure, not the construct, needs modification, then we cannot attempt any construct validation of implicit attitude measures.

This is a prohibitively high cost to pay. For many psychologists, like Messick (1989), it is reason enough to reject the syntactic view. Moreover, as I argued in chapter 2, turning construct validity into an ‘ideal’ that we strive towards but never achieve does not actually solve the problem of validating measures. Psychologists need some account of construct validity that they can actually use in practice.

Our other alternative is to adopt Conventionalism, which amounts to embracing the necessary stipulations involved in measurement validation and providing a descriptive account that systematizes these assumptions Conventionalists, like Mach (1886) and Poincaré (1898), offer up accounts of these stipulations, or principles of coordination. These principles of coordination are conventionally chosen principles used to apply quantity concepts to methods of measurement, like simplicity and uniformity. Principles of coordination, while not empirically verifiable, explain what otherwise would seem like arbitrary decisions made by scientists.
Reichenbach’s example of a coordinative definition: “a measuring rod retains its length when transported” (1958, p. 16) is one such convention. This is a generalization that cannot be empirically verified, and thus according to the verification criterion of meaning is meaningless. Rather than taking such generalizations to be meaningless, Reichenbach argued this statement was an example of an arbitrary rule for determining whether different instances of length are equal. We chose to take the generalization as true in order to maintain consistency across instances of length.

Conventionalism might be a satisfactory solution to the coordination problem if one is committed to hypothesis-testing and the syntactic view. However, the particular issue here with bootstrapping is how one reasons about theory and/or measure revision. Conventionalism provides principles that may explain the initial operationalizations of a measure, but these principles are ill-suited for arbitrating discordant data.

This is the bigger worry about the syntactic view that hypothesis-testing assumes: the failure of the syntactic view to accommodate theory revision. That the syntactic view cannot adequately accommodate or explain theory revision is a complaint that has received extended attention from philosophers of science (see Craver, 2002, for an overview). Kuhn (1962) and Feyerabend (1965) characterize theory revision under the syntactic view as successional, in which revision happens only when there is a paradoxical choice among incommensurable theories. Theory revision, on these accounts, involves a rejection of the old theory and complete replacement with a new theory. Darden (1991; 1977) and Loyd (1988) both argue that the syntactic view neglects gradual and partial types of theory testing and revision. Darden (1991) in particular highlights an aspect of theory revision relevant to the current example: the piecemeal construct, evaluation, and revision, what he calls “reiterative refinement”. Weschler’s refinement of “intelligence” is not a case in which there were two incommensurable theories of intelligence. It is, instead, a case in which one small aspect of the
theory is refined: the link between the construct (intelligence) and a criterion (life-success). If Weschler’s theory revision was the product of incommensurable theories, we would see much wider sweeping changes in how he defined intelligence, and would not expect him to continue to use test items used to measure the previous ‘intelligence’ construct at all.

In summary, hypothesis-testing commits one to a syntactic view of scientific theories. However, the syntactic view has difficulty accounting for psychological theories, which do not contain law-like generalizations. The syntactic view also has problems accounting for theory revision. Cronbach and Meehl describe a bootstrapping reasoning process that requires revising the initially held theory of the construct to be measured. Yet, they provide no account of how one makes determinations about whether discordant data is evidence against one’s theory or one’s measurement. The syntactic view is similarly unable to provide guidance for this reasoning, even if one endorses Conventionalism.

4.3 Limitations for Hypothesis-Testing Beyond the Syntactic View

Setting aside the syntactic view’s problem accommodating theory revision, hypothesis-testing provides an incomplete picture of measurement revision. The view focuses solely on top-down revision while neglecting bottom-up revision. Hypothesis-testing describes construct validation as essentially and only a process of testing hypothesis with a to-be-validated measure. Any revision of theory or measure is then driven by the theory. Evidence of the measure’s validity is relevant in virtue of its relevance to some hypothesis about the target construct. This picture neglects the very important role of bottom-up, or theory-independent, measure revision.
This distinction between top-down and bottom-up validation refers to the difference between revising a measure based on a test of some hypothesis derived from a construct’s theory (top-down), and revising a measure based on either technological advances in a measurement method or evaluations of features of the measure itself, independent of any consideration of the theory of the construct (bottom-up). In top-down measure revision, a particular hypothesis guides the entire experimental process. Evidence is evaluated and then applied to a measure’s validation relative to that hypothesis. In contrast, bottom-up measurement revision is independent of any hypothesis about the construct in question. Technological advances—like the creation of more precise measurement methods, or statistical analysis of the measure itself—guide revision of the measure without any direct consideration of the construct’s theory.

According to hypothesis-testing, the primary source of evidence for a measure’s validity is data produced by the measure that confirms a hypothesis about the measured construct. When a measure fails to confirm a hypothesis, this is evidence either that the measure or theory needs revision. As I discussed in the previous section, Cronbach and Meehl (1955) and syntactic views of scientific theories, have a problem explaining theory revision. While they acknowledge that failure to confirm a hypothesis is evidence for either measure or theory revision, Cronbach and Meehl have no systematic criteria or method for choosing.

Even if we adopt an account of theory revision that provided such criteria, hypothesis-testing also neglects any evidence for construct validity that is not derived from theory testing. Bottom-up construct validation uses evidence not derived from theory testing. Common forms of evidence used in psychological construct validation that I consider bottom-up are technological advances in measurement, development of new statistical methods, or evaluations of features of the measure itself. These are all ways in which evidence is created.
not only absent of any direct test of a theoretical hypothesis about a construct but is also evaluated and used to guide measurement revision independent of any such theory.

Technological advances can lead to a revision of already existing measurement methods or the creation of entirely new measures. The development of a computer program like e-prime that allows researchers to both present stimuli and record response data lead to repurposing old priming paradigms for use on computers. While the conceptual basis of the measure remained the same, the computer-based versions of the tasks were more accurate in recording subjects’ response times. Computer-based tasks could also handle much larger stimuli sets that could be automatically randomized. It is intuitive to believe these computer priming tasks have higher construct validity than their pen and paper predecessors, yet this improvement is driven entirely by advances in technology and not by hypothesis-testing. The revision process here goes in the opposite direction to hypothesis-testing: technology causes measure revision, new measure produces new evidence, new evidence then is applied to one’s theory.

Psychologists also use a range of statistical tests to assess features of tests themselves, independent of any theoretical consideration for the construct the test measures. Test-retest reliability, for example, is a common measure of a test’s reliability over time. Test scores from the same subjects, taken at different times, are correlated. There are field-wide standards that specify what range of scores on these statistical tests is acceptable for a measure. Similar to the standard that p values must be below .05 in order for a difference to be considered statistically significant when applicable test-retest reliability should be above .7.

While reliability is a feature of a measure that can be assessed relatively independent of the construct’s theory, not all statistical tests are equally ‘bottom-up.’ Predictive validity, which refers to the degree to which test scores correlate with a current or future criterion,
can only be assessed if one assumes a theory of the construct that associates the construct with criteria. Predictive validity, then, is assessed only through theory-testing.

Cronbach and Meehl were not unaware of any of these types of evidence. However, they considered these types of evidence to be relevant to other forms of measure validity, independent of construct validity itself. As Messick (1989) argues, this ignores the fact that all these other forms of measure validity are themselves evidence for construct validity. Whether or not a test is reliable is relevant to whether or not that test has construct validity relative to a construct. As what Cronbach and Meehl take to be separate forms of measure validity are all features of a measure that directly limit how construct valid that measure can be, they are, in fact, evidence for a measure’s construct validity. This is why Messick considers construct validity to be the ultimate form of measurement validity (1989, 1995a, 1995b).

Statistical tests are used not only to evaluate measures, but the creation of new statistical methods can also itself cause a drastic revision of both measures and constructs. A prime example of this was the application of factor analysis to intelligence testing, which lead to Spearman’s theory of generalized intelligence, and revision of intelligence test scoring procedures. As I will argue, the use of factor analysis by Spearman to assess the degree of convergence between disparate intelligence measures lead to radical bottom-up theory and measure revision.

4.3.1 Spearman’s G

Factor analysis is a method of statistical analysis that describes variation between observed variables in terms of a potentially lower number of factors, which are unobserved variables. One way to understand factor analysis is it attempts to answer the question: what is the smallest number of constructs (unobserved variables) do can we posit in order to explain the
relationship between x number of observed variables? The motivating assumption for factor analysis is that we can use information about the dependencies between observed variables to reduce the number of variables in a data set. If, for example, we have four domain-specific tests of specialized intelligence (mathematic intelligence, $I_m$, emotional intelligence, $I_e$, creative intelligence, $I_c$, analytical intelligence, $I_i$), each test measures a specific observable intelligence variable ($I_m$, $I_e$, $I_c$, $I_i$). Factor analysis takes the correlations between these variables, and if successful describes the variation in these variables with a smaller number of factors, or even a single factor. A factor analysis of domain-specific tests of specialized intelligence could potentially yield a single factor, $I$, that could predict individual variation on all four tests.

Factor analysis, and related statistical methods like principle component analysis, offer up a way in which statistical methods when applied to measures can drive bottom-up construct revision. One can see how such construct revision would occur with factor analysis in domains in which constructs are hierarchically organized. By hierarchically organized, I mean constructs which are sub-capacities, or instances of, some higher categorical construct. “Memory” is one such higher categorical construct, that encompasses a wide array of sub-capacities like “working memory”, “episodic memory”, “semantic memory”, or “implicit memory”. A factor analysis of each sub-capacity as an individual variable could support revisions to the organization of the different sub-capacities by showing how variation in all sub-capacities, or a subset of the sub-capacities, can be predicted by appealing to a single factor.

English psychologist Charles Spearman (1863-1945) used factor analysis to revise intelligence measurement, and eventually develop his own theory of intelligence. Spearman created and pioneered many statistical analyses in psychology, including Spearman’s rank correlation coefficient, a nonparametric measure of rank correlation. Spearman’s study of intelligence (1904) is regarded as the “earliest version of a 'factor analysis’” (Lovie & Lovie, 1996, p. 81).
Psychologists studying intelligence found themselves in dire straits at the beginning of the 20th century. Correlational analysis, a statistical method by which one estimates the relationship between two or more variables, was still a new statistical method. In the hands of intelligence researchers, correlational analysis had failed to find correlations between a variety of intelligence measures and criterion that researchers initially predicted would be related. As I discussed in chapter one, Wissler (1901) published the most extensive study of potential relationships between intelligence scores and a long list of criteria predicted to be related to intelligence. Yet, even Wissler’s study failed to find significant correlations.

Spearman’s study of intelligence was motivated by the repeated failure of correlational studies of intelligence. Surveying the numerous studies that failed to find statistically significant relationships between intelligence tests and any observable criterion, Spearman remarks:

There is scarcely on positive conclusion concerning the correlation between mental tests and independent practical estimates that has not been with equal force flatly contradicted; and amid this discordance, there is a continually waxing inclination—especially noticeable among the most capable workers and exact results—absolutely to deny any such correlation at all. (1904, p 219)

At the time of Spearman’s study, there was little evidence to justify a general intelligence construct, i.e., some general intelligence capacity that was at least partially responsible for an individual’s performance across different intelligence tests. The repeated failures to find meaningful correlations by psychologists could be considered evidence itself against the hypothesis that there was some general intelligence capacity that these intelligence tests measured. However, Spearman interpreted the lack of observed relationships between supposed measures of intelligence and criteria as a methodological failure, rather than counter evidence to general intelligence. He notes that psychologists rarely used the available correlational measures to produce precise evaluations of their observational data. Moreover, psychologists
failed to assess the accuracy of their measures, in order to estimate the proportion of observed scores that was due to measurement error. This failure was particularly troubling for Spearman, as the methods frequently used to measure intelligence were far from accurate and subject to many different sources of bias. Spearman’s goal, then, was not just to investigate potential relations between intelligence and observable criteria, but to “remedy this deficiency of scientific correlation” by utilizing a new and powerful statistical tool (1904, p. 203).

In his study, Spearman tested male and female school-aged children, all from the same small village in New England, using four different types of intelligence tests: sensory discrimination tasks (as described in chapter 1), school ranking, Teacher assessment, and interview. School ranking is the rank of the child in relation to the rest of their class, corrected for age. Teacher assessment was ranking of each child, brightest to least brightest, by their instructor, not based on any specified criteria other than a general judgment of the child’s intellectual ability. Spearman describes the interview measure as an estimation of “common sense” (p. 251). Spearman interviewed the two oldest children in his sample about their judgments of their peers, based on the criterion of “sharpness and common sense out of school” (p. 251). Each child’s ranked list was solicited independently and compared against a similar list created by the village Rector’s wife (p. 251). Based on the general congruence between the three lists, Spearman judged the children’s peer assessment to be reliable.

Spearman’s employment of factor analysis was the utilization of a new measurement tool—a statistical method—that provided new information about the intelligence construct. His factor analysis yielded the result that variation amongst individual intelligence tests could be reduced down to a single factor:

As the reader will have noticed, the formulae given at the end of the previous chapter are equations whereby from several observed correlations we are able to
deduce a single true one. This latter alone is of real scientific significance. . . (1904, p. 256)

Spearman used $g$ to refer to general intelligence, a shared factor across any test of intellectual ability. $S$, on the other hand, refers to specialized intelligence, a factor unique to the specific sub-ability that a particular intelligence test measured. Thus, according to Spearman, an individual’s performance on an intelligence test could be explained by appealing to the individual’s $s$ and $g$ factors:

Under certain conditions the score of a person at a mental test can be divided into two factors, one of which is always the same in all tests, whereas the other varies from one test to another; the former is called the general factor or G, whilst the other is called the specific factor. (Deary, Lawn, Bartholomen, 2008, p. 9)

Thus Spearman’s development of a statistical analytic method, when applied to the already existing intelligence measures, revealed a new effect that Spearman interpreted as evidence for a general-intelligence construct. As I will argue in the next section, the development of factor analysis is a type of measurement revision undertaken independently of any theoretical assumptions about a particular psychological construct. Therefore, Spearman’s interpretation of his experimental data is an example of bottom-up construct/measure revision.

4.3.2 The Case for Bottom-up Revision

In order to fully understand how Spearman’s factor analysis is indeed a case of bottom-up theory revision, we must distinguish between the specific effects that were revealed by Spearman’s statistical analysis of his experimental data and his later theoretical interpretation of the data as evidence for his particular theory of intelligence. Bogen and Woodward (1988) argue that experimental data are distinct from scientific phenomena. While data is information
created through research that scientists analyze to find effects, phenomena are the stable processes out in the world that scientists attempt to explain via data. Phenomena, on their account, are not directly observable, and can only be inferred from stable effects. Facts about data and facts about phenomena serve as evidence for different types of claims. Whereas facts about data are evidence for claims about phenomena, facts about phenomena themselves are evidence for claims about general theories that attempt to explain the phenomena (p. 306).

Crucially, Bogen and Woodward argue that the stability of effects can, in principle, be judged independent of theoretical considerations (p. 310). In other words, a statistical analysis of data that reveals an effect can be performed without having a particular theory that explains or predicts the effect. The issues and questions that arise about factor analysis, for example, have to do with data analysis and statistical inference, not theoretical questions about a particular construct. There are different methods one can use for factor extraction, the first phase of exploratory factor analysis. Which method one chooses is based on mathematical features of the data, not the theory.

In Spearman’s study, the data he produced were scores on four different types of intelligence tests. The effects he observes in the data are: first, an ordering of scores of the individual students such that scores on each measure varied between individuals; second, that a ranked list of subjects based on their scores was similar across measures; and third, a single factor could statistically explain the similarity across these ranked lists.

The statistical analysis Spearman conducted required no presuppositions about the existence or nature of generalized intelligence. This is not to say that Spearman had no beliefs about intelligence. His motivation to perform that factor analysis was in part a belief that there may be some shared capacity across intelligence tasks. However, the method of statistical analysis was not dependent on this theoretical assumption. If we imagine an alternative
Woodward (2000a) draws a useful distinction between data production and data interpretation. Data production refers to the mechanical process by which a researcher uses a measurement method in an experimental set-up to create effects (p. S165). Data interpretation, in contrast, refers to the process of using statistical analysis and patterns of reasoning in order to draw inferences from the effects to claims about the phenomenon of interest (p. S165).

Woodward makes this distinction in order to argue that the reliability of a measurement is derivable independent of a particular scientific theory. This is possible in cases where one is investigating the “qualitative error characteristics” of the phenomenon in question (p. S169) by focussing on “…a large number of highly specific local empirical facts about the causal characteristics of the detection or measurement process” (p. S170). In these cases, the reliability of the measurement procedure rests not on the more general scientific theory, but rather on empirical facts about the behavior of the measure (p. S172). In other words, a data production process can be judged as reliable independent of a theory when, and if, data interpretation depends only on the empirical results of the experiment, and not on the theory of the phenomenon. As evidence for this claim, he provides examples of instances in science in which scientists established a measurement method was reliable even though the theory they held of the quantity they measured was false. Galileo, for example, developed many empirical arguments for the reliability of his telescope despite the fact that his lacked knowledge of a correct optical theory that would explain how the telescope worked (p. S176).
His claim may be extended here to Spearman’s factor analysis. Woodward’s distinction when applied to Spearman’s study indicates that the factor analysis in Spearman’s experiments changed his data interpretation, rather than his data production, process. The causal process by which Spearman produced his initial data was quite similar to the methods employed by previous psychologists. What Spearman changed was his statistical analysis of the data, attempting to both estimate and correct for measurement error and specify the degree of relation between his measurement methods. Spearman’s statistical analysis is a revision of already existing measurement methods with a novel data interpretation technique.

Understood this way, factor analysis is itself a part of Spearman’s data interpretation. Spearman’s data interpretation requires no derivations of hypothesis about intelligence. The revision, in this case, occurred independently of any particular theoretical assumptions Spearman had about the nature of intelligence. When the revised measurement method produced a new effect—shared variance across data sets attributed to a single statistical factor—Spearman interpreted this as evidence for a general intelligence capacity that was partially responsible for subjects’ performances across different intelligence measures. What we have then, is a case in which a statistical advancement drove a theory revision of a construct in a measurement method. In other words, bottom-up construct revision.

### 4.4 Conclusion

Hypothesis-testing is a method of construct validation rooted in the syntactic view of scientific theories. As such, the ultimate failure of hypothesis-testing is in part due to the inadequacies of the syntactic view. The primary inadequacy here, I have argued, is not necessarily the commitment to describing scientific theories in terms of an axiomatic system. Rather, hypothesis-testing, like the syntactic view, gives an incomplete picture of theory revision.
In particular, I have highlighted here that while hypothesis-testing may, arguably, allow for theory revision when theory-testing fails to confirm a hypothesis, it cannot explain theory revision that is driven by advances to measurement that are theoretically independent.
Chapter 5

Epistemic Iteration

5.1 Introduction

The previous two chapters looked at epistemic theories of justification that psychologists have already utilized. This chapter argues for an alternative theory, *epistemic iteration*, that has not yet been applied to psychological measurement. Epistemic iteration is a view first put forth by Chang (2007) in his work on the history of thermometry. Van Fraassen (2008) advocates for a similar account of measurement more generally in the physical sciences. Epistemic iteration utilizes a coherentist theory of justification. Coherentism, broadly construed, is the view that beliefs are not justified in isolation, but instead, are justified as a system of beliefs. A belief is more or less justified depending on how well the belief agrees with or supports other beliefs in the system (Foley, 1998).

While I will argue here that epistemic iteration is the best epistemic theory of construct validation, it is not without detractors. The primary limitation of Chang’s theory is that without adopting a stricter account of theory coherence, one runs the risk of overly permissive
pluralism about what counts as measurement validation. Therefore, successful application of epistemic iteration to psychology requires additional criteria for evaluating theory coherence beyond what Chang provides.

In section one, I describe the epistemic iteration view of construct validation, focusing primarily on Chang’s (2004) formulation. Chang argues that measures are validated over time through a process of iterative revision, in which new revisions of measures and construct definitions build on proceeding versions. Each new version of either a measure or construct is so revised in order to achieve specific epistemic goals (p. 45). These revisions are justified when they increase the coherence of the theory with these epistemic goals. I then show how this epistemic iteration process unfolds in intelligence testing and how this account of measurement validation permits measure driven theory revision. Thus epistemic iteration accounts for the type of bottom-up revision that hypothesis-testing could not.

This example however also demonstrates a possible limitation of Chang’s account: its reliance on a vague notion of epistemic progress. Chang’s view grounds the justification of a measurement method in an evaluation of epistemic goals. Epistemic goals, on his account, can be an increase of any epistemic virtue, like explanatory power, that are generally considered by philosophers of science to be indicative of scientific progress. His pluralism about epistemic virtues while descriptively accurate raises the possibility that epistemic iteration has no prescriptive weight. I argue that this is a pressing worry for psychologists, as even a single epistemic virtue, like unification, is not applied in the same way across all fields of psychology. I look at neuroscientific research on both implicit memory and intelligence and show how each field has dealt with similar neurological evidence in very different ways.

Epistemic iteration explains the different interpretations of neurological evidence between each fields by appealing to different epistemic goals: neural localization and psychometric
Neural localization is the goal of unifying psychological and neurological theories of a construct by identifying specific neural regions, networks, or anatomical features that are responsible for the construct’s function. Psychometric stability, on the other hand, refers to unification between psychological theories of a construct and psychometric effects, like stable and robust correlations between a construct and observable criterion. In intelligence research, neural localization may require a drastic revision of the psychological theory, and come at the cost of psychometric stability. Thus, while both fields have attempted to unify their psychological theories with neurological evidence, different epistemic goals motivate very different interpretations of the evidence in intelligence research. This raises the question of if and how one can evaluate epistemic goals, especially in cases where these goals conflict. I conclude that my descriptive claims about epistemic virtues in psychology reveal the need for a theory of coherence which constrains potential epistemic goals, giving prescriptive weight to Chang’s view.

In the last section, I provide one example of a theory that could play this role, Thagard’s explanatory coherence theory. His theory provides a set of principles that guide deliberation between conflicting sets of beliefs. Coopting his theory for the evaluation of epistemic goals then introduces a method of deliberating between conflicting epistemic goals. I demonstrate how such deliberation could operate in intelligence research, where the goals of neural localization and psychometric stability currently conflict.

Measures then are justified not by their logical relationship to a construct (operationalism) or their evidentiary relationship to a construct’s theory (hypothesis testing). Instead, they are justified by their relationship to the epistemic goals of a research project. The implications of my extension of epistemic iteration to psychological construct validation are that successful validation requires endorsing both some account of coherence and knowledge of the specific epistemic goals of one’s research domain. This permits a tempered pluralism about epistemic
goals in psychology while maintaining criteria for deliberation between conflicting epistemic goals. For philosophers, my argument here identifies the need for an analysis of psychological research in terms of epistemic goals against a theory of coherence.

5.2 Coherentism about Measurement

Coherentism about measurement validation is a view advanced by Chang with his account of *epistemic iteration* (2001, 2004, 2007, 2016a, 2016b) and by van Fraassen in his more recent defenses of *constructive empiricism* (2008, 2009, 2012). For coherentist epistemic theories, a belief is justified by the relationship between that belief and the system of beliefs it belongs to. In contrast, foundationalist epistemologies determine the justification of a belief by evaluating its relationship with other, foundational, beliefs. Thus coherentist epistemologies determine justification holistically, while foundationalist epistemologies determine justification by appealing to a special class of self-justifying beliefs. “Coherentists deny that any beliefs are self-justifying and propose instead that beliefs are justified in so far as they belong to a system of beliefs that are mutually supportive.” (Foley, 1998, p. 158-159).

When applied to measurement validation, coherentism leads to a radically different solution to the coordination problem. This approach takes a historical perspective on measure-construct relationships in order to demonstrate how definitions of constructs and their measures are justified over time, in tandem. Whereas operationalism attempted to avoid the circularity of the coordination problem altogether, and conventionalists propose principles of coordination that can ground coordination, coherentists embrace the circularity and attempt to show that it is not vicious.

My focus here is on Chang’s (2004) formulation of coherentism in his explanation of the development and validation of thermometers through a process of *epistemic iteration*. Chang’s
epistemic iteration theory of measurement makes three key claims: 1. Measures must be evaluated from a historical perspective, and from that perspective coordination between measures and constructs ceases to be viciously circular, 2. That measurement validation is a process of epistemic iteration, and 3. A measurement method is justified if it increases the coherence of the theory compared to previous iterations. I will first explain Chang’s view, along with van Fraassen’s argument that historical perspectives on measurement escape the coordination problem. I will then return to the previous chapter’s example of Spearman’s general intelligence theory, in order to demonstrate how epistemic iteration works.

I conclude, however, that Chang’s account is incomplete. As epistemic iteration relies on evaluations of scientific progress in order to determine whether a measure is justified or not, Chang’s gesture at potential epistemic virtues is inadequate. While he is correct that there is surface level agreement about likely candidates for epistemic virtues of scientific theories, there is substantial disagreement between accounts of individual virtues. A reductive account of theory unification, for example, conflicts with non-reductive accounts of the ‘same’ virtue. Consequently, what it means to achieve theory unification greatly depends on which account you endorse. Further, epistemic goals falling under the umbrella of a single virtue, like unification, can conflict. I will present an example of such conflict in the following section. Thus, the successful application of epistemic iteration to psychology requires further philosophical work on how to deliberate between potentially conflicting epistemic goals that arise from different accounts of epistemic virtues.

5.2.1 Historical Perspectives on the Coordination Problem

Chang (2004) and van Fraassen (2008) argue that the validity of a measure must be evaluated from a historical perspective, in which the entire history of both the construct and measure is taken into account. This solution is different from both operationalism and hypothesis-testing,
which attempt to evaluate measures in historical isolation, i.e., relative only to the current evidence for the measure’s interpretation and theory of the target construct. Both operationalism and hypothesis-testing attempt to be perspective neutral about justification, taking what van Fraassen calls a “view from nowhere” (2008, p. 122). According to operationalism, the WAIS-IV is valid only in virtue of whether it is a successful operationalization of intelligence. According to hypothesis-Testing, the WAIS-IV is valid in virtue of the test’s ability to confirm hypotheses derived from the current theory of intelligence. Previous versions of the WAIS have no bearing on the validity of the current version; justification is an ahistorical evaluative claim. Instead, coherentists evaluate a measure’s validity by taking into account both the current evidence and the relationship between the measure and previous measurement methods that the measure builds upon. For example, whether or not the current (IV) version of the WAIS (Weschler Adult Intelligence Scale) is a good measure of intelligence is a question answered by both evidence about the current test and previous versions of the test. That, comparatively, the WAIS-IV may do certain things better than the WAIS-III is relevant to the validity of the WAIS-IV. A measure’s validity is a relative judgment, made by comparison between a measure and what came before it.

Due to this shift to a historical perspective, coherentism provides a much different solution to the coordination problem. Recall that the coordination problem is that the questions, “What is X?”, and “How do you measure X?” circularly presuppose answers to one another. As both operationalism and hypothesis-testing assume an ahistorical perspective, their solutions to the coordination problem amount to assuming the answer to “What is X?” in order to answer “How do you measure X?”. However, as discussed in the previous two chapters, this is an unsatisfying solution as it leads to viciously circular reasoning. A measurement method is justified because it has the right relationship with a stipulated definition of the target
construct. Neither operationalism or hypothesis-testing provides an adequate account of how a construct’s definition moves beyond stipulation.

The major shift for coherentism is not to deny the need for stipulation, but rather view any stipulation as a temporary assumption that is part of a larger historical trajectory. Coordination is not achieved by fixing one question and then answering the other. Instead, coordination is achieved through a process of give and take. We temporarily stipulate a definition of a construct in order to develop a measure, which in turn produces evidence used to revise the definition.

Operationalism and hypothesis-testing describe measurement validation as a uni-directional process, in which there is a fixed starting point: a theoretical definition of the construct. For operationalism, measurement validation is merely a process of deducing the operationalization of the construct entailed by that definition. I have argued that this fails to absolve the vicious circularity of measurement validation because the initial stipulation is never justified. Hypothesis-testing loosens the relationship between construct and measure; however like operationalism, it stills fails to provide an account of how this stipulation is justified. Even further, while hypothesis-testing does not in-principle deny theory revision, I argued in the previous chapter that the account of revision here is incomplete. Hypothesis-testing only accounts for revision driven by empirical tests of hypothesis about the construct, neglecting revisions driven by technological advances in measurement. Moreover, hypothesis-testing provides no principles for determining when failed tests should lead to measure revision or theory revision.

Chang and van Fraassen argue that from a historical perspective, coordination between construct and measure ceases to be viscously circular:
The coordination process between measure and construct ceases to be vicious when viewed “from above”, i.e., retrospectively with our current scientific knowledge, or “from within”, i.e., examining advances in their original, historical context. (Van Fraassen, 2008, p. 122)

The stipulation required for coordination is temporarily justified by its instrumental use in developing an initial measurement method. This stipulation is non-vicious because it is corrected or revised in the future by the very measurement procedure it enables.

Chang permits, albeit temporary, foundational beliefs. His view is then not strictly coherentist, as foundational beliefs are self-justifying. For this reason, he refers to his view as “progressive coherentism” (2007). Chang writes:

This doctrine is coherentist in the sense that it rejects the search for an absolutely firm foundation. At the same time, it recognizes the possibility of making progress by first accepting a system of knowledge without ultimate justification, and then using that system to launch lines of inquiry which can, in the end, refine and correct the initially affirmed system. (p. 5)

In summary, coherentists embrace the circularity in measurement validation and attempt to show that it is not vicious. They do so by arguing that from a historical perspective, measure and construct are validated in tandem. Temporary stipulations that are instrumentally necessary in order to develop an initial measurement method do not lead to viciously circular reasoning because these stipulations are eventually revised, or corrected, by the research program they made possible.

5.2.2 Epistemic Iteration

Chang describes the in tandem coordination of measures and constructs as a process of \textit{epistemic iteration}. Using the history of thermometry, Chang argues that the construction
of a quantity-concept and the standardization of its measurement are iterative processes that depend on each other. With each advance of scientists’ technical ability to measure temperature, there is a new iteration of the theoretical definition of temperature that reflects this technological advancement. At the same time, new measurement methods allow scientists to collect new data, which leads to revision in their theory of the construct. The theory of a construct is modified as its measurement procedure changes and vice versa. This account embraces the circularity of coordination while rejecting its viciousness. Measure and construct are coordinated, not due to unjustified stipulations, but due to this ongoing, iterative process.

This process is iterative in that measurement methods and constructs go through multiple revisions. It is epistemic in the sense that each iteration is developed in order to advance certain epistemic goals. The epistemic goals researchers work towards vary based on the particular stage of development they are engaged in.

Chang’s epistemic iteration approach to construct validation consists of three stages. First, there is an acceptance of an existing system of knowledge that enables the development of a research program, including a measurement method. Second, this measurement method is used to investigate the target construct. In this stage, researchers attempt to refine the measurement method in order to achieve sufficient reliability and precision. Third, the refined measurement method is used to develop the theory of the target construct further.

Prior to the first stage, there is no scientific measurement procedure, save for our senses. Constructs in the first stage are folk properties that are measurable only in a “crude and limited sense” (p. 87) through direct observation. In the case of thermometry, this stage consisted of human experiences of hot and cold. Temperature is “measured” by our tactile experience of heat. Our sensations of temperature can differentiate between hot and cold, but we struggle to make fine-grained distinctions and are prone to error in our judgments.
under extreme circumstances. For example, running one’s hand under room temperature water after you have been outside in the cold can cause one to judge the water to be scalding hot erroneously.

The first epistemic goal for researchers is to develop a measurement method that enables scientific investigation of a particular phenomenon of interest. This first stage of epistemic iteration is only possible when researchers accept some existing theory about the phenomenon, even though the theory may lack justification (p. 6). The acceptance of a prior standard is necessary because it gives researchers guidance for the creation of a measurement procedure that, while grounded by sensory experience, can build on that experience. The principle of respect refers to the non-absolute nature of this acceptance: “...our use of instruments is made with a respect for sensation as a prior standard, but that does not mean that the verdict of sensation has unconditional authority.” (p. 43).

Once a measurement method is created, researchers can test hypotheses (given by the initially adopted theory) about a construct and revise both their theory and measurement procedure. Crucially, observations now aided by the new measurement procedure may conflict with unaided observations. For example, the thermoscope, the predecessor to the modern thermometer, enabled measurements that revealed observations of temperature that were finer grained and more consistent than raw sensory experience.

Thermoscopes measure changes in temperature through the rise and fall of a liquid contained in a tube. Initial thermoscopes were developed based on the prior sensory standard. As one sensed a substance, like a bowl of water, became hotter to the touch, the liquid in the thermoscope would rise. It was this agreement between thermoscopes and sensory experience of heat that justified their initial use.
Once markings were added to thermoscopes that enabled more precise observations of changes in the level of the liquid, hypotheses could be tested. Researchers observed that a thermoscope would report the same temperature in a liquid even when two individuals’ tactile experience of the liquid differed. Thus, the initial measurement procedure is first grounded by sensory experience. Though iterative refinement a measurement procedure is developed that then augments or even corrects our sensory experiences of the construct.

The second stage of epistemic iteration seeks to achieve a measurement method that is reliable and precise enough to enable numerical assignments to the quantity, i.e., construct, being measured. A reliable measure, or measures, will give the same reading in identical situations, whereas a precise measure will be able to differentiate between different situations.

Chang refers to the effort to increase the reliability and precision of a measure as a criterion of comparability (p. 89). The epistemic goal for researchers is to be able to compare between measurement methods in the same experimental setting, and across different experimental settings with the same measurement method.

Chang (2004) argues that researchers achieve this goal only by adopting the principle of single value. The principle of single value is that “a physical property can have no more than one definite value in a given situation” (p. 90; see Ellis, 1969, p. 39). From this principle, it follows that a measure should indicate only one definite value in a given situation. Comparability then requires that measurement procedures are reliable (i.e., reproducible) and precise (i.e., free from random error) enough that the measure can differentiate between different situations and be consistent in identical situations.

The third stage of epistemic iteration is reached once a numerical standard has been established (p. 87). Once a numerical standard is established, researchers can more precisely test hypotheses about the construct, working towards more general epistemic goals.
The epistemic iteration Change speaks of happens throughout the second and third stages. As hypotheses are tested, the results lead to changes in the working theory of the construct, which results in a new iteration of the construct. These theory revisions, in turn, leads to new hypotheses to be tested. With each new iteration of a construct or measurement method, researchers attempt to improve on the previous version in order to move closer to their epistemic goals.

What epistemic goals are, for Chang, is related to epistemic virtues. Roughly, any epistemic goal of a research field are subgoals of the more general goal to increase the epistemic virtues of theory. The goal of reaching a numerical standard, then, is a subgoal of the more general goal of increasing the explanatory power of one’s theory. Epistemic virtues for Chang are any feature of a theory that could plausibly be considered to be indicative of scientific progress. A prime example would be explanatory power. Hempel et al. (1965) defines explanatory power as the ability of a theory to explain, such that if theory, T₁, explains more phenomena than T₂, T₁ is a better theory than T₂ because it has more explanatory power.

Progress is achieved, and thus measures are justified, when iterations increase coherence between the theory/measure and an epistemic goal. In other words, if a revision increases some epistemic virtue, it increases coherence and is therefore justified. Chang is markedly pluralistic in what he thinks denotes increased coherence:

Whether and to what extent an iterative procedure has resulted in progress can be judged by seeing whether the system of knowledge has improved in any of its epistemic virtues. Here I am defining progress in a pluralistic way: the enhancement of any feature that is generally recognized as an epistemic virtue constitutes progress. (Chang, 2004, p. 228)
For many, this appeal to general epistemic virtues may be surprising due to the existing and ongoing disagreement about epistemic virtues in philosophy of science. Chang (2004, ch: 5) argues that despite ongoing philosophical debates that would indicate otherwise, there is a surprising amount of agreement among philosophers about what are epistemic virtues.

He is right, to a point. Hempel (1966, p. 33-46) gave criteria of confirmation and acceptability of a theory: quantity, variety, and precision of supporting evidence; support by more general theories; ability to predict previously unknown phenomena; simplicity; and credibility relative to background knowledge. Kuhn (1977, p. 322) discussed “standard criteria for evaluating the adequacy of a theory”: accuracy, consistency, scope, simplicity, and fruitfulness. Van Fraassen (1980, p. 87) referred to pragmatic virtues (which on his account justify a scientific theory): elegance, simplicity, completeness, unifying power, and explanatory power. Lycan (1988, 1998, p. 341) provides the following epistemic, or “theoretical” virtues: simplicity, testability, fertility, neatness, conservativeness, and generality of explanatory power. Chang, however, sees a significant degree of overlap, despite profound philosophical disagreement between these philosophers about scientific progress. He gives the following four virtues are a rough approximation of what most philosophers of science take to be epistemic virtues: 1. Empirical success, 2. Explanatory power, 3. Unification 4. Simplicity/Parsimony (p. 229).

I will return to this issue of defining coherence in the next section, as whether or not I can construct a plausible theory of coherence is directly relevant to epistemic iteration’s success as a theory of psychological construct validation. Before I turn to possible limitations of epistemic iteration, I want to demonstrate why epistemic iteration is the most descriptively accurate theory of measurement validation examined so far.
5.2.3 Epistemic Iteration in Psychology

On Chang’s account, epistemic iteration is a process of coordinating measure and construct in tandem. The primary benefit of this account is that it characterizes the revisionary relationship between measure and construct as genuinely bi-directional. A theory of a construct and the hypotheses it entails may drive revisions of a measure. At the same time, advances in measures, including those that are not theory-driven but technologically driven, can drive revisions of the theory. This account accommodates the bottom-up theory revision that hypothesis-Testing neglects. Therefore, epistemic iteration, unlike hypothesis-testing, or operationalism for that matter, accounts for all forms of measure/theory revision.

Epistemic iteration is best understood by way of an example. We can apply Chang’s epistemic iteration to my example of test development and revision of intelligence testing described in previous chapters. I will show here how his account is, unlike the previous epistemic theories considered, able to accommodate measure driven theory revision.

Spearman’s application of factor analysis to intelligence testing is a prime example of measure-driven construct revision. Recall that Spearman (1904) used a new statistical analysis, factor analysis, to analyze data from various already existing intelligence measurement methods. The result was that variance across these measures could be explained by a single factor, what Spearman called $g$ or general intelligence. Spearman developed a theory of general intelligence that proposed a general intelligence capacity, $g$, that is utilized in every intelligence task, and a specialized intelligence capacity, $s$, which is utilized only in a specific kind of task. Thus an individual’s performance on an intelligence test is explained by appealing to their $g$ and their $s$ relevant to the specific task the test taps in to.
According to Chang, the way to understand this episode is as follows. First, there is the pre-Spearman construct of intelligence, in which intelligence is defined as some general problem-solving capacity. Along with this, there are a variety of intelligence measures, which included sensory discrimination tasks, peer judgments, researcher interviews, etc. The primary problem with these measures was that they had poor predictive, criterion, and even convergent validity. Individual measures, like sensory discrimination tasks, were poor predictors of future criterion, like educational attainment (Wissler, 1901). They also had low, and often insignificant, correlations with concurrent criteria, like school grades. Finally, intelligence measures had low correlations with each other spearman1904.

Due to the problems with intelligence tests, intelligence theory was problematically incoherent. There was little agreement between the individual measures which did not correlate, and intelligence theory, that predicted agreement, i.e., coherence, among intelligence tests. Another way to put this is intelligence theory posited the existence of a general capacity that could be measured by intelligence tests. As each test was thought to measure the same capacity, the theory predicted some agreement across the tests. However, research pre-Spearman contradicted this prediction.

Spearman’s factor analysis is, as I argued in the previous chapter, a technological advancement in measurement. The mathematical development of factor analysis, when applied to already existing measures of intelligence, created a new method of analyzing data. Thus, factor analysis leads to a new iteration of intelligence testing.

Crucially, whether or not this new iteration of intelligence testing yielded evidence for a hypothesized general intelligence capacity did not drive the measure revision. One way to see this is to think of the alternative scenario in which Spearman’s factor analysis failed to find an underlying factor. It is unlikely that Spearman would have taken this as evidence to
revise his method of factor analysis, as the justification for and development of this statistical analysis was mathematical, and not contingent on any particular psychological theory. While factor analysis was used to test a particular hypothesis about intelligence, the initial revision was caused by advances in mathematics unrelated to advances in psychological theories.

This new iteration of intelligence testing yielded new evidence—primarily the shared variance across different intelligence measures—that then justified Spearman’s revision of intelligence theory. Spearman’s theory of general and specialized intelligence was then a new iteration of the intelligence construct, revised in response to the advancement in measurement.

Further, the theoretical prediction Spearman tested is an example of the criterion of comparability applied to a psychological construct. If intelligence is, in fact, a psychological quantity, it cannot, according to the principle of single value, have more than one value at a specific time in an individual. Thus tests that measure intelligence must be consistent between each other when given to the same subject. The efforts of Spearman, and other psychologists, to develop psychological measures of intelligence that correlated with one another is an example of the second stage of epistemic iteration.

The final piece of the epistemic iteration story is the evaluation of the new iterations of both measure and construct. These iterations are justified if they increase coherence among the theory. While I give a more in-depth analysis of coherence in the next section, for our present purposes a rough working definition will suffice. Coherence is a property of the relations between elements of a theory, like theoretical definitions or models of constructs, hypotheses, and empirical observations. These elements cohere to the degree that they ‘hang together’ rather than conflict. A paradigmatic example of coherence in this context would be evidence that confirms a prediction derived from a theory, whereas evidence that disconfirms a prediction would be incoherent with the theory.
There are two relations in intelligence theory that were improved by Spearman’s revision: the relationship between measures and the construct, and the relationship between the theory’s predictions and data. First, Spearman’s theory not only revealed convergence between intelligence measures but also provided a method of evaluating individual intelligence measures so that they could be improved upon. Spearman was able to show that there was, in fact, concordance between intelligence measures. Moreover, \( g \)-factor, which is calculated from multiple intelligence measures, could, in turn, be used to evaluate individual intelligence measures. The correlation between a specific test and \( g \) is then an approximation of the degree to which the test measure generalized, rather than specialized, intelligence. This is now referred to as the \( g \)-loading of a test. Thus, Spearman’s revision improved the relationship between measures and construct such that it was possible to determine how well a particular measure captured the construct, and even further revise the measure to improve the relationship.

Second, Spearman’s analysis led to new evidence that better-supported predictions about intelligence. The factor analysis demonstrated convergence across intelligence measures, supporting the prediction that if each test measures intelligence, the test should agree with other measures of intelligence. \( G \)-factor is also a better predictor of a variety of criterion—concurrent and future—that intelligence theory claimed were related to intelligence.

Given that Spearman’s new theory of intelligence increased agreement between theoretical predictions and measures, and amongst measures themselves, coherentists would conclude that it is a justified iteration.
5.3 Evaluating Epistemic Iteration

In the previous chapters, I have argued that both operationalism and hypothesis-testing fail as theories of psychological construct validation because they 1. do not solve the coordination problem, and 2. cannot handle psychological constructs. Epistemic iteration, in contrast, does both.

5.3.1 De-fanging the Coordination Problem

Epistemic iteration solves the coordination problem by embracing the circularity involved in measurement validation and demonstrating how it need not be vicious. Construct validation does require initial theoretical stipulations during the creation of a new measure. However, Chang (2004) and van Fraassen (2008) argue that these stipulations are temporarily justified by their instrumental value, and revised later on when the research project yields new empirical observations. Thus, as long as psychology only assumes what is necessary to develop a measure, and allows future revision of these assumptions, we need not worry that coordination is viciously circular. As van Fraassen writes:

*There is no presuppositionless starting point for coordination.* We are, to adapt Neurath’s metaphor, sailors engaged in a continuing construction, renovation, alteration, and repair of the ways in which we measure our ship while already at sea. (2008, p. 137)

It is worth belaboring Neurath’s metaphor here. In order to set sail, we are forced to use the less-than-desirable wood to fix the ship’s hull. However, once the ship is sea-worthy, we can then sail to a better island where we can find sturdier wood.
The historical examples I have examined throughout this dissertation involve the very kind of revisions that epistemic iteration allows, including revisions to initial assumptions about the nature of intelligence. The wide-scale rejection of sensorimotor discrimination tasks after Wissler’s (1901) study of the correlation between such tasks and academic performance is one such revision. The initial assumption that there existed a relationship between sensorimotor discrimination and intelligence enabled the development of early intelligence tests. These measures were then used to test this relationship. When these tests failed, psychologists revised their theory and rejected those measures as non-valid measures of intelligence.

The eventual empirical observations almost a century later that there did exist a small, negative correlation between sensorimotor discrimination and intelligence (A. R. Jensen, 1982) is not evidence of faulty reasoning by psychologists about intelligence measures. Rather, it further demonstrates their willingness to revise previously held theoretical beliefs about the validity of their measures based on new evidence.

The solution offered by epistemic iteration amounts to shifting our perspective on the reasoning process involved in measurement validation. This perspective shift is most evident when contrasting epistemic iteration with the hypothesis-testing approach described in chapter 4. Recall that hypothesis-testing relies on a syntactic view of scientific theories, such that construct validation is a top-down, deductive process. On this view, construct validity depends wholly on the empirical observations produced by a measurement method and their evidentiary relationship to hypotheses derived from the theory of the target construct. Construct validation, then, involves reasoning about the relation between empirical observations and theory.

Advocates of the syntactic view of scientific theories examine scientific theories in isolation, as linguistic structures that follow a logical framework. Whether or not an empirical observation
supports a hypothesis, and therefore is evidence for the construct validity of a measure, is determine solely by the scientific theory. Chang and van Fraassen argue that the coordination of theoretical terms and measurement procedures only appears to involve viciously circular reasoning when we restrict ourselves to the limited perspective of the syntactic view. Coordination appears to be circular as what justifies the measurement method are the very hypotheses about the target construct we are testing.

Epistemic iteration, in contrast, broadens the scope of the reasoning involved in construct validation in order to include the broader historical context and the epistemic goals of the researchers. The construct validity of a measure cannot be determined solely from the evidence that measure produces for hypotheses about the target construct. Instead, previous measures, and the evidence they produced for earlier iterations of the relevant theory (what van Fraassen refers to as “prior achievements” (2008, p. 124)), are relevant to the validity of the current measure. These prior achievements enable investigations that produce new evidence, which in turn enables the theory to advance. At each step of the coordination process, the iteration of the theory and measure is justified by its relation to epistemic goals and previous iterations. Construct validation involves reasoning then about the relation between the current measure, previous measures, and the epistemic goals of the researchers. By broadening the scope of reasoning, epistemic iteration eliminates the necessity for circular reasoning.

5.3.2 Accommodating Psychological Constructs

Further, Chang’s account of epistemic iteration can accommodate psychological constructs, be they lumpy, under-defined, or unstable. Measurement validation does not require a nomological network containing the construct in question, nor does it require exhaustive operationalization. All that is needed is an initial theory of the construct that supports a
measurement method. Thus early intelligence researchers did not need to know whether intelligence was stable across one’s lifetime, or that it was related to cognitive capacities like working memory or attention. All that was need was the thought that there existed some general cognitive capacity predicted an individual’s performance across a variety of problem-solving tasks, and a willingness to revise their initial assumptions.

Cronbach & Meehl talked of the need for psychologists to “lift themselves up by their bootstraps”, yet were unable to explain how such bootstrapping was possible. When measures produce discordant data, Hypothesis-testing provides no guidance about choosing between revising one’s measures or revising one’s theory. Epistemic iteration is the account of bootstrapping that Cronbach & Meehl could not provide. Researchers can choose between measure and theory revision based on the particular epistemic goals of the current research project. Determining whether an iteration of either a measure or a theory is justified depends on whether it advances these epistemic goals. The consequence of this account is that determining whether or not a measure is justified requires first knowing what the particular epistemic goals are, and some account of coherence that enables evaluating success in reaching those goals.

5.4 Coherence and Epistemic Goals

The previous section’s example of epistemic iteration in intelligence testing begs the question, what counts as increasing coherence between measure and theory?

As I remarked in the previous section, Chang is pluralistic about epistemic virtues, and by connection, coherence. His use of coherence is not as an epistemic belief justification theory, per se, but rather as a theory of goal achievement. He views coherence as a feature of the relationship between scientific activities and epistemic goals, such that an activity
is coherent with a goal if it advances it. The target of his coherentism is not beliefs, then, but scientific activities and epistemic goals. Thus an epistemic iteration of a measure is an activity, of revising one’s measure and/or theory, and the epistemic goals refer to increasing any epistemic virtue of the theory:

Whether and to what extent an iterative procedure has resulted in progress can be judged by seeing whether the system of knowledge has improved in any of its epistemic virtues. Here I am defining progress in a pluralistic way: the enhancement of any feature that is generally recognized as an epistemic virtue constitutes progress. [p. 228]chang2004

On Chang’s account, whether or not a measure is valid depends then on whether the iteration of the measure increased an epistemic virtue of the relevant scientific theory.

In the previous section, I outlined four general epistemic virtues that Chang references: 1. Empirical success, 2. Explanatory power, 3. Unification 4. Simplicity/Parsimony (2004, p. 229). The epistemic goals that Chang attaches to stages of epistemic iteration can be thought of as specific goals that enable the advancement of these more general epistemic virtues. For example, the standardization of a scale that enables comparisons across contexts and measurement methods is a precondition for increasing the empirical success of a theory. Scientists must be able to quantify the phenomenon in order to test a theory’s predictions. Quantification requires a measure with a standardized scale that can measure whether a phenomenon is the same or different across manipulations. Testing a theory’s predictions, when successful, increases the theory’s empirical success. Thus, scale standardization is a specific epistemic goal that advances the more general epistemic goal of increasing a theory’s empirical success.
Chang’s gesture at epistemic virtues like parsimony and unification is unsatisfying when one considers the plurality of accounts available for individual epistemic virtues. If coherence is measured by an increase in any of these epistemic virtues, coherence judgments require choosing between multiple, sometimes conflicting, definitions. There is a real worry here that coherence defined so pluralistically becomes useless for making any prescriptive claims about measurement validation. If a revision of a measure does fit with one account, there are many others to chose from.

In this section, I will examine two potential problems with Chang’s account that arise from his commitment to coherentism. First, his pluralism about epistemic virtues, and as a consequence, coherence, leads to the worry that coherentism has no prescriptive weight. If any revision counts as scientific progress, we should worry that Chang’s formulation of coherentism fails to enable the basic comparative claims between measures needed for an account of construct validation. This worry is particularly relevant to psychological research, which epistemic goals vary between and even within subfields. Applying epistemic iteration to psychology requires some account of how to deliberate between conflicting epistemic goals.

Second, with any coherence theory of justification comes the possibility that our theories end up spinning in a void, without corresponding, or connecting with, reality in any meaningful way. The worry for any coherence theory of justification is the threat of endorsing coherent theories that end up being false. A set of statements can be logically consistent but contain false statements. Likewise, we should worry that a measurement method can be coherent with a theory of a construct even though our theory of that construct is false. Thagard (2007) summarizes the problem as follows:

A major issue for coherentist epistemology concerns whether we are ever warranted in concluding that the most coherent account is true. In the philosophy of science,
the problem of coherence and truth is part of the ongoing controversy about scientific realism, the view that science aims at and to some extent succeeds in achieving true theories. The history of science is replete with highly coherent theories that have turned out to be false, which may suggest that coherence with empirical evidence is a poor guide to truth. (p. 28)

In order for epistemic iteration to be a successful theory of psychological construct validation, we need some account of coherence that has perscriptilve weight, and avoids the worry of coherent-but-false theories. Here, I will look at two different examples from research psychology that demonstrate how epistemic goals may conflict. Implicit memory and intelligence researchers have attempted to unify psychological theories of their construct with neurological theories by identifying potential neural correlates, i.e., brain regions that realize the construct’s function. However, the drastically different interpretations of similar neurological evidence in both sub-fields indicate that implicit memory researchers have different epistemic goals than intelligence researchers. Thus the different notions of unification at play in psychological are more than a philosophical issue; it leads to drastically different epistemic goals and as a result different interpretations of neurological evidence that conflicts with a psychological theory.

Neuroscientific investigations of intelligence demonstrate this problem. Intelligence researchers have attempted to identify a “neural g” that would unify psychological theories of intelligence with neuroscientific theories. However, this neural localization goal conflicts with another unification goal: psychometric stability. Successful neural localization may require drastic revision of both psychological theories of intelligence and intelligence tests. Researchers want to keep intelligence tests in their current form, which produce stable, robust, and statistically significant effects—for example, a high correlation between intelligence and educational achievement (Deary et al., 2007). Thus, the goal of neural localization conflicts with the goal of psychometric stability.
5.4.1 Varieties of Unification

Chang’s pluralism about epistemic virtues becomes problematic once we consider specific, mutually exclusive accounts of a single virtue. Take, for example, unification. Unity accounts can be either ontological or epistemic (and sometimes both). Ontological accounts of unity focus on the metaphysical aspects of unifying relations between elements of theories/models. Epistemic accounts focus instead on epistemic relations between elements of theories/models and epistemic goals like the explanation of phenomena.

The strongest formulations of unification are reductive accounts. In reductive unification, two theories are unified when either one theory is reduced to another theory, or both are subsumed under a more general theory. An example of the classic definition of reductive unification comes from Nagel (1961). On his account, theory T_1 is reduced to theory T_2 iff the laws of T_1 can be derived from the laws of T_2 by way of bridge principles. Bridge principles are biconditionals that link terms in the vocabularies of each theory. Reductive unification originates in syntactic views of scientific theories, discussed in the previous chapter. As such, reductive unification takes the unification of two theories to require the right sort of logical relationship between the laws of the two theories. The laws of the reduced theory must be derivable from the laws of the theory to which it is reduced.

Connective unification encompasses any account of unification that does not require reduction. In this way, connective unification is a weaker notion than reductive unification, as these types of accounts do not require deductive logical relations between theories.

While some philosophers still endorse a form of reductive unification (e.g., Bickle, 2003, 2006, 2007; Churchland, 1982), the general consensus is that especially in the “special sciences” like psychology and neuroscience, connective unification is more likely. Even
reductive accounts, like Bickle’s, are not a classic intertheoretical reduction account. On Bickle’s account, low-level areas of neuroscience redefine psychological constructs in terms of observable behavioral changes which are then accounted for by molecular mechanisms. Thus, psychological constructs are redefined and then explained by neuroscientific theories. However, there are no psychological laws that are derivable from the laws of neuroscience.

A classic account of connective unification comes from Darden & Maull (1977). Their account looks at inter-field theories, theories that describe relations between scientific fields. Darden & Maull characterize scientific fields as areas of science that attempt to solve the same scientific problem(s) using shared researcher methods. For example, neuroscientific research of long-term potentiation (LTP) and cognitive psychological research of memory investigates the same phenomena (memory) but using different measurement methods and research techniques. Any theory that unifies findings from each field would be an inter-field theory.

Darden & Maull argue that two fields using different research methods to look at similar phenomena posit theories that may be connected, even though the theories often do not compete and are not logically reducible to one another. A new field, like functional neuroimaging, can specify the location, structure, function, and causal relevance of processes posited in a different field, like cognitive psychology. However neuroscientific theories can be connected to psychological theories without reducing or redefining psychological constructs in neuroscientific terms.

A related perspective on unification comes from mechanistic theories of scientific explanation (Baetu, 2011; Bechtel, 1986, 2006; Bickle et al., 1998; Bickle, 2006; Craver, 2007; Craver & Darden, 2013). Accounts of mechanistic unification are derived from mechanistic theories of scientific explanation. Briefly, mechanistic accounts of explanation posit that an explanation explains some phenomena if it describes the mechanism that causes the phenomena.
Mechanistic philosophers of science understand fields of science through mechanism schemas. Unification occurs when one field constrains the organization of, entities within, or processes of mechanisms that explain a common phenomenon from another field (Craver, 2007). However, mechanistic unification is not reductive. Each field contributes a part of the mechanistic explanation of a phenomenon, including psychology. Psychological explanations are not reduced to neurological explanations.

The brief overview here clearly shows that theory unification is not one, unified epistemic virtue. Rather, philosophers define unification in a variety of ways, each of which has different consequences for science-in-practice. However, the multiplicity of unification is not in of itself problematic. If it is the case that this multiplicity is merely a philosophical issue, rather than a scientific one, then the problem is that philosophers have not correctly identified the theory of theory unification. Coherentism could have prescriptive weight, once philosophers do the leg work of ’unifying’ theory unification.

That is not the case. Surveying sub-fields of psychology reveals that the multiplicity of unification is more than a philosophical issue. Scientists, especially in psychology, in practice, have different pragmatic objectives that lead them to evaluate unity with respect do different aims. This multiplicity is particularly evident if one looks at the various ways psychologists reason about neuroimaging studies of psychological constructs.

5.4.2 Unification in Psychology

Theory unification in psychology is often interpreted as theory unification between psychological and neuroscientific theories. This requires then accommodating neuroscientific evidence in psychological theories of cognitive constructs, like implicit memory or intelligence. One way in which psychologists incorporate neuroscientific evidence into their research of psychological
phenomena is neural localization. Neural localization refers to the practice of identifying the neural correlates of a psychological construct, a neurological feature that is thought to be, in part, causally responsible for the construct. In other words, these neural correlates are the physical, neurological mechanisms that produce the psychological phenomena associated with the construct.

The success of neural localization varies widely in psychology. Whether or not it is even in principle possible is an open question for both neuroscientists and philosophers alike (see Milkowski, 2016). There are two primary questions up for debate: 1. What neural features are best candidates for neural correlates? and 2. What is the correct relationship between a psychological construct and neural correlate?. These are questions asked both globally across all of cognitive science and within specific research fields. These are not purely philosophical questions; they have real consequences for how researchers reason about neuroscientific evidence. Neural localization is not applied consistently across, or even within, research projects in psychology. Thus, researchers themselves disagree about the answers to both questions.

One problem is that neuroimaging data is often discordant with psychological constructs. The way psychologists have carved up cognitive abilities does not correspond, in any direct fashion, with the categories supported by neurological evidence. When evidence is discordant, it is an open question for psychologists whether this justifies revising the measurement methods or the psychological theory of the construct. Alternatively, psychologists can ignore discordant neurological evidence, and privilege evidence from other domains.

Here, I will briefly consider two examples of how psychologists accommodate neuroimaging data that conflicts with their constructs. In some fields, like implicit memory, neural evidence is used to justify drastic theory revisions such that the initial features that distinguished
between constructs are abandoned. Psychologists replace them with features that better align with neural correlates. In other fields, like intelligence testing, neural evidence is taken as tangential to other justifications for the construct, like stable psychometric effects.

One way to understand the differences in these two cases is to attribute different epistemic goals to each sub-field. While implicit memory and intelligence researchers are interpreting similar evidence, they reach different conclusions due to the particular epistemic goal of their sub-field. This reveals, however, that a ‘single’ epistemic virtue, like “unification” does not directly translate into a universal epistemic goal.

### 5.4.3 Implicit Memory

Implicit memory, as discussed briefly in chapter 2, is for some a problematic cognitive capacity due to the lack of any plausible unique psychological mechanism, or even neural mechanism. Roediger (2003) argued that implicit memory’s theoretical definition is so broad that it encompasses all biological phenomena, not just psychological phenomena. Moreover, there seems to be no cognitive process that unifies all instances of implicit memory.

Willingham & Preuss (1995) argue that not only is there no unique cognitive mechanism for implicit memory, there is also no unique neurological mechanism. In other words, there are no neurological correlates to implicit memory. Like Roediger, Willingham & Preuss conclude that implicit memory is not a legitimate construct; there is no implicit memory capacity that is distinct from other forms of memory. We see here, then, a convergence between some memory researchers that neural localization of implicit memory has failed, and that failure signals major problems with the implicit memory construct.

Recall that implicit memory was initially distinguished as a distinct memory capacity from explicit memory due to single-dissociation evidence from amnesic patients. Patients with
damage to their medial temporal lobes (MTL) are impaired in explicit memory retrieval tasks, but unimpaired in implicit memory tasks (Carlesimo et al., 1999; Eichenbaum & Cohen, 2001; Squire & Zola-Morgan, 1991; Squire, Stark, & Clark, 2004). Willingham & Preuss argue that single dissociations that demonstrate the lack of involvement of specific neural structures in implicit memory are not sufficient. In order for the term implicit memory to be legitimate, there needs to be anatomical consistency across implicit memory itself (1995, p. 2). On this front, implicit memory fails. Impairment on implicit memory tasks due to brain damage is inconsistent across measurement procedures. Patients with brain damage have impaired performance on some implicit memory tasks and not others. For example, Heindel, Salmon, Shults, Walicke, and Butters (1989), and Harrington, Haaland, Yeo, and Marder (1990) demonstrated double dissociations between motor skill learning and other implicit tasks in patients with damage to their striatum due to Huntington’s disease. Parkinson patients similarly were impaired in their ability to learn the pursuit rotor motor skill. However, Huntington’s patients had normal repetition priming, and Parkinson’s patients learned the mirror reading perceptual skill normally (Harrington et al., 1990). From this, Warrington & Preuss conclude:

No one has proposed any manner in which to tie together the apparent neural separability of these different tasks. That is, one might argue that the implicit memory system is, in fact, a coherent system. Although it is subserved by a number of structures, those structures are interconnected, and to the extent that one sees dissociations with the system, that simply reflects the differential effects of damage to isolated components. The system, nevertheless, is coherent. Such an argument is theoretically possible, but it has not been made, to our knowledge, nor does one seem likely. (p. 4)

While there seems to be a general consensus that implicit and explicit memory depend on distinct brain systems (Tulving & Schacter, 1990), Warrington & Preuss rightly point out that there is a difference between showing that implicit memory utilizes different brain systems
than explicit memory, and showing that there is anatomical consistency in the brain systems utilizes across all instances of implicit memory. The former is evidence merely for drawing a distinction, while the latter is the evidence needed to justify unifying implicit memory as a cohesive cognitive capacity.

More recent studies provide evidence that there is not anatomical consistency across all instances of implicit memory. The brain regions utilized in implicit memory studies vary depending on the type of stimuli and type of implicit memory task. Perhaps unsurprisingly, modality of stimuli determines which brain regions are employed. Neural priming is observed in the occipital cortex for visually perceived stimuli (Badgaiyan & Posner, 1996; Badgaiyan, 2000; Zago, Fenske, Aminoff, & Bar, 2005), the fusiform cortex for objects or face stimuli (Henson, Shallice, & Dolan, 2000), and the left inferior prefrontal cortex for lexical or semantic stimuli (Buckner, Koutstaal, Schacter, & Rosen, 2000; Wagner, Paré-Blagoev, Clark, & Poldrack, 2001).

In fact, the only consistency in neural activation across different priming tasks is that neural activity decreases as stimuli are repeated (Henson et al., 2000; Berry, Shanks, & Henson, 2008; Soldan, Gazes, Hilton, & Stern, 2008). Thus while there are clear differences between explicit and implicit memory tasks in terms of which neural regions are involved, and the direction of neural activity, there is/are no clear neural region(s) that unifies all instances of implicit memory or that are unique to implicit memory.

Accommodating the heterogeneous evidence of potential neural correlates for implicit memory lead Dew and Cabeza (2011) to propose a radical revision of how memory researchers distinguish between implicit and explicit memory. Their theory identifies five neural regions associated with memory (the hippocampus, left ventrolateral prefrontal cortex, rhinal cortex, parahippocampal cortex, and visual cortex) that are related to three features of a memory
process: 1. The cognitive process (whether it is conceptually or perceptually driven), 2. The stimulus representation (item or relational), and 3. The level of intention (controlled or automatic/involuntary). They give the following model:
Figure 5.1

Figure 4. This model predicts that the brain regions associated with explicit or implicit memory do not divide on consciousness, but rather vary along the continua of several critical variables, including the cognitive process (conceptually or perceptually driven), the stimulus representation (item or relational), and the level of intention (controlled or automatic/involuntary). The model de-emphasizes the traditional systems view and leaves open the opportunity for neural regions to contribute uniquely as well as work in synchrony to support various memory phenomena. Hipp, hippocampus; IVLPFC, left ventrolateral prefrontal cortex; RhC, rhinal cortex; PHC, parahippocampal cortex; Vis Ctx, visual cortex.
As the model illustrates, the amount of involvement of a particular neural region is predicted by where the specific memory task falls along the continuum of these three features. A visual priming task, like a word-stem completion task, is cognitive, perceptual, involves item rather than relational memory, and is automatic. Their model then predicts that the brain region(s) most likely operative during the task is the visual cortex and to a lesser degree the rhinal cortex and the parahippocampal cortex.

One consequence of this theory is that there is no clear delineation between explicit and implicit memory or even different forms of implicit memory, that is grounded in a specific neuroanatomical feature. Instead, explicit and implicit memory are cognitive processes that utilize a shared set of functionally connected neural regions. Specific instances of each memory then differ in the degree to which they utilize these different neural regions, depending on features of the encoding/retrieval process (unintentional/intentional), and features of the information encoded (perceptual/conceptual; relational/item).

In summary, the psychological construct “implicit memory” is commonly distinguished as a memory capacity based the retrieval process involved. However, efforts to identify the neural correlates to implicit memory have produced evidence that this theoretical definition of implicit memory does not correspond neatly to any specific neural networks or regions. In response to this discordance between the psychological theory of implicit memory and neural evidence, some have argued for radical revisions to psychological theory. Here I have given an example of one such revisionary theory, Dew & Cabeza’s (2011) theory in which implicit memory is distinguished from explicit memory based not merely on the intentionality of the retrieval process, but also the stimulus representation, and the cognitive process involved.
5.4.4 Intelligence

As with implicit memory, neuroscientific studies of intelligence produce evidence that is potentially incongruous with psychological theories of the construct. No theory is able to explain all the available evidence, and doing so would require radical revision of both intelligence tests and intelligence theory.

For example, the Parietal-Frontal Integration Theory (P-FIT) locates intelligence in a network that includes the parietal and frontal brain regions, along with the white matter structures that link those regions (R. Jung & RJ, 2007). While they derive their theory from a metanalysis of more than 30 neuroimaging studies, few of the discrete brain regions included in the model approach even 50% convergence (Colom, Jung, & Haier, 2007). Moreover, not one of the proposed brain regions is identified across neuroimaging methodologies (ibid.).

Even more troubling is the fact that evidence for the P-FIT model is inconsistent between different intelligence tests. Jung & Haier (2009) found that correlations between gray matter and g-scores depended in part on the tests used to calculate g. Recall that g, is measured by averaging a subject’s intelligence scores across multiple tests. Thus Jung & Haier’s investigation of their theory found that the strength of association between the regions they claim are the neural correlates of intelligence depends on what set of intelligence tests are used to calculate g.

Furthermore, Colom et al. (2006) found that the higher a test’s g-loading, the more widespread discrete brain areas are activated, including many areas not included in the P-FIT model. ‘G-loading’ is a statistical measure of the degree to which a particular intelligence test measures g rather than other, possibly confounding, mental processes. It is calculated by correlating scores on an individual test to the g factor calculated by average across multiple tests. The
higher the correlation, the higher that test’s g-loading, and the more it is thought measure intelligence. This evidence is difficult for any theory that attempts to localize intelligence to discrete brain regions, or even networks, to explain.

For some, the failure to neural localization suggests that discrete brain regions are not even the right targets. Blair (2007), for example, argues:

It may also suggest that there are no specific cortical areas that underlie intelligence, but that individual differences in intelligence reflect aspects of brain function that enable more efficient use of cortical structures and resources that are associated with specific cognitive abilities. (p. 154).

While many criticize neural theories of intelligence, like P-FIT, few intelligence researchers place blame with how general intelligence is defined, or general intelligence theory. Euler (2018) writes that, “…despite this progress [in neural localization], theoretical accounts of these phenomena leave many unanswered questions, and often only indirectly inform debates about the nature of intelligence” (p. 94). In fact, most of the commentary by intelligence researchers focuses on issues of measurement. See, for example, Haier’s (2009) study in which they claim that heterogeneity of neural evidence across studies is due to inadequate methods of calculating g. See also Colom (2007), who argues that neuroimaging studies need to “refine the way this construct is measured” by developing better test batteries (p. 156).

Further, criticism of intelligence research continues to focus not on the discrepancies with the neuroimaging evidence, but problems with intelligence tests’ predictive ability in educational settings (Benson, 2003). These critiques admonish intelligence tests for either being culturally biased or for failing to predict educational outcomes for certain populations (Kaufman, 2007). They criticize the lack of the convergence of intelligence test scores with observable

---

7G-loading is an example of a direct statistical measure of a test’s content validity.
criteria in sub-populations, not the lack of convergence between intelligence test scores and neuroimaging.

Lastly, those that continue to develop neural theories of intelligence have moved away from models that identify specific brain regions, and towards theories that identify more global neuroanatomical features. The Neural Efficiency Hypothesis is one such model, that links intelligence to white matter structure (Li et al., 2009; Penke et al., 2012; Schmithorst, Wilke, Dardzinski, & Holland, 2005) and neural metabolite concentrations (Aydin, Uysal, Yakut, Emiroglu, & Yılmaz, 2012; R. E. Jung et al., 1999, 2009; Nikolaidis et al., 2016). These neural theories attempt to explain both the neuroimaging and psychometric evidence without revising psychological theories of intelligence.

The psychometric evidence for intelligence are robust population-level effects. Individual differences in intelligence are normally distributed in the population (Deary, Penke, & Johnson, 2010). Your intelligence measured at 11 will correlate highly with your intelligence measured at 79, as will a rank ordering of you and your peers based on intelligence test scores (Deary, Whalley, Lemmon, Crawford, & Starr, 2000). Intelligence tests have high predictive power, for education achievement (Deary et al., 2007), occupational attainment and social mobility strenze2007, and job performance (L. S. Gottfredson, 1997). In regards to these psychometric effects, Deary et al. (2010) write, ”Individual differences in human intelligence are among the most robust observations in psychology” (p. 201).

In summary, intelligence researchers are willing to revise just about anything except the theory of intelligence itself. Rather than revise their construct, they propose revising their measures or revising their neural theories. They do so in order to maintain the predictive power of intelligence tests, along with the robust effects observed in the population. Their
goal seems to be maintaining psychometric stability, i.e., continuing to use intelligence tests that produce robust population-level effects, rather than neural localization.

5.4.5 Conflicting Epistemic Goals

Coherentism tells us to evaluate measures respective to epistemic goals. Measures that move us closer to these goals are justified. However, how researchers interpret general epistemic virtues, like unification, into more specific goals that can be achieved in the lab, is context dependent. From this context dependencies, we get varied, and even conflicting goals that promote drastically different interpretations of similar empirical evidence.

Both of these examples show psychologists dealing with similar forms of neuroimaging evidence. Neuroimaging evidence is incongruent with the demarcated psychological construct, such that the construct does not neatly align with a specific brain region or network. The persistent problem of interpreting discordant data rears its head again. The question is, do psychologists revise their measures, so that they produce data congruent with the neuroimaging data, or do they revise their theory so that the construct aligns with the neuroimaging data?

Implicit memory researchers seem to favor modifying their theory. Intelligence researchers, on the other hand, take a third option: ignore the neuroimaging data in order to keep theory, and measure, aligned with a different set of predictive criteria.

Implicit memory researchers, like Willingham & Preuss (1995), emphasize the need for neuroanatomical consistency for memory capacities, such that discrete brain regions and/or networks are responsible for a capacities function. Thus, revisionary models of implicit memory, like Dew & Cabez’s (2011), are developed with the intention of increasing neuroanatomical consistency. Intelligence researchers, on the other hand, emphasize the need to respect the psychometric properties of intelligence, i.e., robust evidentiary patterns in intelligence scores
and their correlations with observable criteria. In order to maintain psychometric stability, intelligence researchers ignore neural evidence that could justify revising their theories of intelligence. Further, while the specific goals of each field differ, both appear to be utilizing some notion of theory unification. Yet, each field unifies their theory with a different type of data.

Chang’s pluralism then is descriptively adequate; however, the worry for psychologists is that epistemic iteration fails to have prescriptive weight. If even singular epistemic virtues like unification can in practice provide multiple different epistemic goals, any revision of a measure or theory can be post hoc justified by identifying the right epistemic goal. Intelligence researchers could equally justify radical revision of their theory in order to achieve neural localization, or ignoring the neural evidence in order to maintain psychometric stability. The more general worry here is that if any epistemic goal will do, one can always find an epistemic goal that would justify a particular measure or theory revision.

5.5 Coherence from the Perspective of Psychologists

This move from pluralism about epistemic goals to denying prescriptive weight is too hasty. I argue that if we are clearer about what we mean by increasing coherence, epistemic iteration can have prescriptive weight. Moreover, the right account of coherence can avoid extreme relativism about truth.

5.5.1 Systems of Practice

Chang’s pluralism about scientific progress reflects his larger commitment to both pluralism and pragmatism about philosophy of science (see Chang, 2007, 2012, 2016a, 2016b). He has argued elsewhere for analyzing scientific knowledge in terms of “epistemic activities” and

In this regard, Chang follows practice-oriented approaches to scientific inquiry. For example, Longino (1990) applies MacIntyre’s (1999) concept of practice to scientific inquiry. According to MacIntyre, a practice is “any coherent and complex form of socially established cooperative human activity through which goods internal to that form of activity are realized in the course of trying to achieve those standards of excellence which are appropriate to, and partly definitive of, that form of activity.” (1997, p. 187) Longino contrasts her approach of analyzing science as practice to the traditional approach of focusing on scientific theories. Scientific theories are not in of themselves inappropriate objects of study for philosophy of science, however Longino argues that “In our fascination with individual theories it is easy to lose sight of the fact that scientific inquiry is a collaborative human activity and consequently to approach the methodology of inquiry with tools for the analysis of theories. Theories, however, are the outcome of inquiry and not the process itself.” (p. 17)

Epistemic iteration take a similar practice-oriented approach to measurement validation. Chang (2012, 2014, 2016b, 2017) proposes that scientific inquiry “can be analyzed in terms of ‘epistemic activities’ and ‘systems of practice’”. Epistemic activities are coherent sets of operations, both mental and physical, that are knowledge-seeking “in a particular way in accordance with some discernible rules” (2016b, p. 7). Systems of practice consist of epistemic activities that interact, and are coherent with, one another (2016b, p. 7).

Measurement is itself just one type of epistemic activity. Other epistemic activities are prediction, hypothesis-testing, even “paper-and-pencil operations” (Bridgman, 1959, p. 3). Whether or not a system of practice is coherent is entirely dependent on the overall aims.

185
For Chang, the relevant notion of coherence here is *pragmatic*. *Pragmatic coherence* is not a feature of merely the propositions in a theory, but rather:

...coherence consists in various activities coming together in an effective way toward the achievement of the aims of the system. Coherence comes in degrees and different shapes, and it is necessarily a less precise concept than consistency, which comes well defined through logical axioms. (Chang, 2016b, p. 8)

A direct consequence of this approach is that when Chang argues for coherentism about measurement, his commitment to coherentism is not about the justification of beliefs, but rather about the relationship between scientific activities and epistemic goals. Scientific activities, then, and not beliefs, are justified to the extent that they further the epistemic goals of a research field. He writes:

...the coherence of a system goes beyond mere consistency between the propositions involved in its activities; rather, coherence consists in various activities coming together in an effective way toward the achievement of the aims of the system. Coherence comes in degrees and different shapes, and it is necessarily a less precise concept than consistency, which comes well defined through logical axioms. (2017, p. 8)

When we understand coherence as a claim about the relationship between scientific activities and epistemic goals, Chang’s pluralism about epistemic virtues is less problematic. It is not the case that any interpretation of any epistemic virtue may be used a coherency criteria against which we judge progress. Rather, “It is the overall aims of a system of practice that define what it means for the system to be coherent.” (p. 7, 2017).
5.5.2 Evaluating Epistemic Goals

Let us return to the example of the contrasting interpretations of conflicting neurological evidence between implicit memory and intelligence research. Chang’s proposal guides us to evaluate each case not relative to some overarching theory of coherence, but rather relative to the aims of the system of practice. Thus rather than determine some neural localization criteria that explains both cases, we need to explicate the aims of each system of practice. Instead of asking, “what makes this system of beliefs more or less coherent?”, we need to identify systems of practice, determine the epistemic goals of these systems. Only then can we determine whether or not a particular epistemic activity, like measurement, is coherent with these goals.

These different interpretations of discordant data in the above examples reflect different systems of practice with different epistemic goals. Within implicit memory research, the motivating epistemic goal is neural localization. Within intelligence research, the motivating epistemic goal is psychometric stability. In each system of practice, the discordant neurological data is interpreted relative to these epistemic goals. Each field responds differently to similar neurological evidence because they have different epistemic goals. Both interpretations of the discordant neural evidence are coherent, relative to these epistemic goals.

We may still worry that while epistemic iteration explains why similar data is interpreted differently across different systems of practice, psychology still needs prescriptive norms about the epistemic goals themselves. Coherence with epistemic goals is too permissive a standard if any epistemic goal is legitimate.

While a comprehensive argument for a particular theory of coherence is beyond the scope of this dissertation, I will utilize one particular theory to show how it can guide our evaluation
of epistemic goals. It may seem that Chang’s account is incompatible with more traditional epistemic theories of coherence, due to his focus on goals and activities rather than beliefs and sets of beliefs. However, this is not necessarily the case. The right account of belief coherence can work in tandem with Chang’s view, in order to constrain the set of epistemic goals from which scientists choose. In this way, epistemic iteration has prescriptive weight, and maintains a degree of pluralism about epistemic goals.

Thagard (2007) argues for a theory of *explanatory coherence* in scientific reasoning “that involves theories that progressively broaden and deepen over time, where broadening is explanation of new facts and deepening is explanation of why the theory works.” (p. 29). His explanatory coherence theory (1989, 1997, 2000, 2007) consists of the following set of principles:

**Principle E1 (Symmetry)** Explanatory coherence is a symmetric relation. That is, two propositions A and B cohere with each other equally.

**Principle E2 (Explanation)**

a A hypothesis coheres with what it explains, which can either be evidence or another hypothesis.

b Hypotheses that together explain some other proposition cohere with each other.

c The more hypotheses it takes to explain something, the lower the degree of coherence.

**Principle E3 (Analogy)** Similar hypotheses that explain similar pieces of evidence cohere.

**Principle E4 (Data Priority)** Propositions that describe the results of observation have a degree of acceptability on their own.

**Principle E5 (Contradiction)** Contradictory propositions are incoherent with each other.


**Principle E6 (Competition)** If A and B both explain a proposition, and if A and B are not explanatorily connected, then A and B are incoherent with each other (A and B are explanatorily connected if one explains the other or if together they explain something).

**Principle E7 (Acceptance)** The acceptability of a proposition in a system of propositions depends on its coherence with them.

Thagard describes these principles as algorithms that enable one to determine whether one should reject or accept two conflicting beliefs. On his view, coherence is defined not as a logical relationship between beliefs, or even a probabilistic relationship, but rather an explanatory relationship. While not necessarily incompatible with alternative interpretations of coherence, Thagard’s view is not a purely coherentist view due to Principle E4 (Data Priority). This principle imports some prioritization for empirical observation in of itself, independent of that empirical observation’s relationship with a larger set of beliefs.

This weak pseudo-foundationalist principle makes explanatory coherence a fitting parallel for Chang’s pragmatic coherentism. Just as Chang permitted temporary theoretical stipulations that are instrumentally justified, Thagard’s view permits prioritizing observations. Data Priority is justified because the principle helps Thagard avoid coherent, but false, theories.

A related worry in philosophy of science is Newton-Smith’s (2002) pessimistic induction: the inference that any scientific theory will eventually be discovered to be false. The history of science is replete with examples that support this inference. One of the most infamous is the phlogiston theory of chemistry, discussed in the second chapter. Phlogiston theory was, at one time, very coherent. It explained chemical phenomena like combustion. Thus explanation alone does not guarantee the truth of a theory.
Thagard’s solution to the pessimistic induction is to say that the best theory is not just the one that best explains the evidence, but also “broadens its evidence base over time and is deepened by explanations of why the theory works.” (p. 37). He refers to this as the “deepening maxim”, and it, along with the Data Priority Principle, aim at empirically grounding scientific theories.

Deepening an explanation amounts to expanding the evidentiary scope of the explanation by introducing a new hypothesis. In science, Thagard argues that deepening usually involves providing an underlying causal basis, i.e., a causal mechanism, for the phenomenon (2007). For example, Dew & Cabeza’s model of implicit memory posits multiple neural regions that are responsible for implicit and explicit memory. The degree of involvement of a particular region is determined by features of the memory task, like the modality of the information being processed. This model deepens previous explanations of implicit memory by explaining more evidence, in particular evidence that neural regions involved in implicit memory are heterogeneous across different information modalities. It also yields new hypotheses about the involvement of specific neural regions. Testing these hypotheses creates more evidence that the model can then, if successful, explain. This further deepens the theory.

Explanatory coherence, as the name implies, relies on a notion of “explanation”. Thagard himself favors mechanistic accounts, briefly outlined earlier in this chapter. A mechanistic account of scientific explanation offers up a useful schema but is not necessary for Thagard’s view. One could plausibly endorse an alternative theory of explanation and still accept his principles.

The utility of Thagard’s theory for our present purposes is that these principles, along with the deepening maxim can also determine whether one rejects or accepts a particular epistemic
goal, rather than a belief. If achieving a goal requires violating one of these principles, then the goal should be rejected.

We can then apply Thagard’s theory to the previous two examples of interpreting neural evidence. On the one hand, implicit memory researchers have the primary goal of neural localization, whereas intelligence researchers have the primary goal of psychometric stability, with neural localization as a secondary goal. The principle of explanatory coherence mandates that we privilege fewer hypotheses to explain the same evidence. In the case of implicit memory, neural localization models like Dew & Cabeza’s do just that, by revising the theory in order to deepen the explanation to include neuroimaging evidence. Note that this does not mean that neural localization is the only epistemic goal relevant to implicit memory research. Rather, Thagard’s theory justifies its use, as it is a goal that increases theory coherence.

Intelligence researchers, on the other hand, have yet to offer up revisions that accommodate all of the relevant neuroimaging evidence. This does not necessarily mean that they are wrong to ignore discordant neuroimaging evidence. Rather, their failure to accommodate neuroimaging evidence must be explained by an epistemic goal in accordance with Thagard’s principles. Psychometric stability does seem to respect these principles, in so far as intelligence theory is able to explain psychometric effects. Moreover, if neuroimaging evidence is discordant with a psychological theory, one could justify that accommodating the evidence comes at the cost of explaining these psychometric effects. If accommodating the neuroimaging evidence requires modifying intelligence tests such that they fail to be diagnostically predictive, one has not really increased the coherence of the theory. One has merely traded one set of explained data for another. However, if there were a theory that could accommodate both psychometric and neurological evidence, Thagard’s principles would mandate acceptance of such a theory, even if it meant radically revising our intelligence tests.
The benefit of applying Thagard’s view in this way is that it can explain how a specific epistemic goal, like psychometric stability, is appropriate relative to a certain evidentiary state of affairs. As it currently stands, psychometric stability is a worthy epistemic goal only if it yields a theory of intelligence that accommodates the most evidence with the fewest hypothesis. This will, and can, change as more evidence is gathered. Thus just as the validity of a measure changes over time with epistemic iteration, the appropriateness of a specific epistemic goals changes in accordance with the available evidence and theory.

5.6 Conclusion

Psychological measurement is a necessary component for psychological research, yet the reasoning involved in the creation and justification of such measures, what psychologists refer to as construct validation, has received little attention from philosophers of science. I have argued that psychological measurement deserves attention because it involves potentially viciously circular reasoning, referred to as the coordination problem. Solutions to other forms of circular reasoning in science are ill-suited to the coordination problem in psychology due to features of psychological research, and psychological constructs.

The solution offered here requires adapting, and extending, coherentist accounts of measurement by Chang & van Fraassen (Chang, 2004, 2007, 2009a, 2014, 2016a, 2016b; Van Fraassen, 2008). Chang’s account of epistemic iteration can accommodate psychological measures and adequately solve the coordination problem. However, developing a coherentist view of construct validation requires developing an account of coherence that can be applied to epistemic goals. Otherwise, pluralism about epistemic virtues, and as a result, epistemic goals, leads to a view with little prescriptive weight.
Through my presentation of two research projects: implicit memory and intelligence testing, I demonstrated that even a specific epistemic virtue like unification operates differently across psychology. In implicit memory research, psychologists have prioritized convergence between neuroimaging data and their psychological theories of implicit memory. This guides them to revise their functional analysis of implicit memory to reflect the constraints placed by neuroscience. In intelligence research, on the other hand, psychologists have prioritized convergence between intelligence test scores and predictive criteria, so that intelligence tests remain useful diagnostic tools across a variety of contexts.

This points to a need for philosophical work, not about the epistemic virtues themselves, but how they apply differentially across fields. In addition, in order for epistemic iteration to have prescriptive strength, philosophers need to go beyond mere description of the notions of epistemic virtues, and by connection, epistemic goals, at play.

I have provided a brief sketch of how one such account of coherence, Thagard’s explanatory coherence theory, could offer constraints for potential epistemic goals. Thagard’s theory proposes a set of criteria for deliberating between two conflicting beliefs. In this context, it enables deliberation between two conflicting epistemic goals: neural localization and psychometric stability.

Thagard’s criteria are especially helpful here, as he focuses on the explanatory relations between beliefs and evidence. When coopted for epistemic goal deliberation, Thagard’s criteria make the deliberation dependent on the particular evidentiary status of the field in question.

The consequence of such an account for psychologists is that measurement validation requires knowledge of the epistemic goals of the particular field, along with an analysis of such goals as the conform or fail to conform to a theory of coherence. My intent here is not to argue
that Thagard’s view is the only such theory that could play this role. Rather, I used his view to demonstrate how a theory of coherence in conjunction with epistemic iteration can prevent overly permissive pluralism about epistemic goals.
5.7 Introduction

I have made two prescriptive claims about psychological construct validity relevant to psychologists. The first is that construct validity should be distinguished from construct legitimacy, as each responds to different evidence. The second is that the construct validity of a measure should be evaluated from a coherentist framework.

5.8 Validity vs Legitimacy

Distinguishing between construct validity and legitimacy means appreciating that things previously considered to be evidence for the validity of a measure are, in fact, directly evidence only for the legitimacy of a construct, and indirectly evidence for a measure’s validity. A prime example of such evidence is evidence that supports a particular model or theory of a psychological construct. Empirical evidence for a particular model of a construct, like implicit attitudes, that supports the model at the exclusion of all other models directly increases the legitimacy of that construct. The same evidence may also increase the construct validity of a measure, but this relationship is indirect and contingent. If evidence converges on a model of implicit attitudes that proposes a underlying causal process by which implicit attitudes...
are formed and retrieved, then it is possible that the construct validity of some implicit attitude measures may increases. This result, however, is contingent on measures existing that tap into the right causal process. It is also contingent on researchers having evidence that the measure(s) tap into the right causal process. One can imagine a situation in which, while researchers agree on a particular model of implicit attitudes, the model singles out a causal process that no current implicit attitude measures tap into. In that case, the construct validity of implicit attitude measures would decrease, even though the legitimacy of the construct has increased. One can also imagine a situation in which we have incomplete knowledge of the causal process(es) involved in our available measures. In that case, the construct validity of the measures would remain the same.

Evidence for the construct validity of a measure is also indirectly and contingently evidence for the legitimacy of a construct. Modifications of a measurement procedure that increase the reliability and/or precision directly increase the measure’s validity. These modifications increase the legitimacy of the measured construct only if they lead to the production of new and better evidence for a theory of the construct. One can imagine situations in which creating better measures results in delegitimizing a construct. The creation of neural measures employed in implicit memory research is one such example. fMRI data provided evidence of heterogeneous neural activation across different implicit memory tasks, despite robust single dissociation effect patterns in the behavior data from the same tasks. . .

The consequence for psychologists is that evaluating evidence for construct validity is context dependent. The same form of evidence will not result in the same increase or decrease of validity in every situation.
5.9 Coherentism

I have argued here that psychological construct validation faces a coordination problem that the current epistemic theories implicitly or explicitly used by psychologists can not solve. I have offered an alternative theory, coherentism, that solves the coordination problem by showing how the problem is not so circularly vicious after all.

Coherentism requires us all, philosophers and psychologists alike, to shift our perspective on construct validation. Rather than evaluate measures using only the current evidence we deem relevant to their validity, like a multi-trait, multi-method matrix, or measures of predictive power, we must also evaluate measures relative to their previous iterations. The question is not then, “Is the WAIS IV a valid measure of intelligence?” Instead, the question is “Is the WAIS IV a better measure compared to the previous version”.

What it means for a measure to be better is determined by the epistemic goals of the research field. This creates a degree of pluralism about construct validation. As particular epistemic goals vary, so will the standards for construct validation. This variation is exemplified by the different responses to incongruent neurological evidence between implicit memory and intelligence researchers. Intelligence researchers endorse an epistemic goal of psychometric stability such that it is more important to maintain robust, population level effects related with intelligence test scores, than to revise intelligence tests to accommodate neurological evidence. Implicit memory researchers, on the other hand, endorse neurological localization, and therefor have proposed drastic revisions of their theory in order to better accommodate the neurological evidence.

There are a few consequences of this account, for both psychologists and philosophers of science.
The main consequence of coherentism is that the standards for construct validity are relative to a specific research project. There are no universal criteria that all psychologists can rely. This is in direct conflict with how some psychologists currently write about construct validity. In his retrospective on the history of construct validity, Smith (2005 focuses on the numerous statistical tests developed with the aim of measuring the construct validity of a measure. The motivation for these statisticians is to create a method of quantifying the evidence for construct validity in a standardized way that can be applied throughout psychology. Coherentism tells us that while such measures may be useful, they will always be incomplete. Correlations between scores, individual test items, and other criteria, only tell us about the test itself. But construct validity is determined by how the test does relative to epistemic goals. This evaluation is not one that can be reduced to statistical measures.

For psychologists, construct validation requires that they pay close attention to the epistemic goals of their research field. These goals determine their standards for construct validity. These goals are often implicit, thus this may require theoretical work by both psychologist and philosophers to explicate what epistemic goals are at play. As I demonstrated in the previous chapter, some epistemic goals are only achieved at the cost of abandoning others. Awareness of one’s goals not only enables construct validation, but also leads to awareness about these potential conflicts. It may be the case that researchers decide that a particular goal is important enough to warrant the cost. Considering the diagnostic applications of intelligence tests in educational, military, and hiring contexts, sacrificing neural localization for psychometric stability may be a justified cost for intelligence researchers. However that cost benefit calculation may, and should, change over time as new evidence is gathered. One can imagine a future in which a generalized intelligence test is developed with the expressed goal of identifying the neural substrates of intelligence, that sacrifices some predictive power in order to do so.
The intelligence testing case exemplifies how researchers privilege certain goals over others even when both goals are thought to advance the same overarching goal (theory unification). Explicating and evaluating epistemic goals is a task that can be greatly aided by philosophers of science. Philosophers of science have always worked to make the implicit assumptions at work in scientific research explicit. Thus, philosophers of science are well situated to do the more global cost-benefit analysis of privileging certain epistemic goals over others. As I demonstrated here, already existing work on coherence from epistemology can be coopted to help with the evaluation of epistemic goals themselves.

Utilizing epistemic theories of coherence, however, raises some old problems from epistemology that philosophers of science must tackle. Every theory of coherence has detractors, and it is up to philosophers to determine which detractors are relevant to the evaluation of epistemic goals in science. Thus while coherentism eliminates one philosophical problem, the coordination problem, it replaces it with another, the deliberation between conflicting epistemic goals.

What remains then is a potential solution to an old problem in measurement that creates a need for cross-disciplinary work from both philosophers and psychologists. Both in the explication of epistemic goals and in the evaluation of and deliberation between these goals.

5.10 Conclusion

Measurement in psychology is still hard as a coherentist. It is just difficult in a different way. Rather that try to search for presupposition-less grounding of our measures, we must strive for a more wholistic understanding of how a single measure fits in to its current theory, aims of the research project, and all the iterations of the measure that came before it. This points to a need for more philosophical work on the epistemic goals in psychology, and how they vary between research projects.
References


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


Cattell, R. B. (1946). Description and measurement of personality.

202


1932 scottish mental survey. *Intelligence, 28*(1), 49–55.


Jung, R. E., Gasparovic, C., Chavez, R. S., Caprihan, A., Barrow, R., & Yeo, R. A. (2009). Imaging intelligence with proton magnetic resonance spectroscopy. *Intelligence, 37*(2),


Thagard, P. (2000). *Coherence in thought and action (life and mind: Philosophical issues in*


Woodward, J. (2002). There is no such thing as a ceteris paribus law. In *Ceterus paribus laws* (pp. 27–52). Springer.
