


## A dialectic on validity: Explanation-focused and the many ways of being human

Bruno D. Zumbo \*

<sup>1</sup>University of British Columbia, Measurement, Evaluation, & Research Methodology Program, Department of Educational and Counselling Psychology, and Special Education

### ARTICLE HISTORY

Received: Dec. 17, 2023

Accepted: Dec. 19, 2023

### Keywords:

Validity,  
Validation,  
Test theory,  
Assessment  
consequences,  
True score.

**Abstract:** In line with the journal volume’s theme, this essay considers lessons from the past and visions for the future of test validity. In the first part of the essay, a description of historical trends in test validity since the early 1900s leads to the natural question of whether the discipline has progressed in its definition and description of test validity. There is no single agreed-upon definition of test validity; however, there is a marked coalescing of explanation-centered views at the meta-level. The second part of the essay focuses on the author’s development of an explanation-focused view of validity theory with aligned validation methods. The confluence of ideas that motivated and influenced the development of a coherent view of test validity as the explanation for the test score variation and validation is the process of developing and testing the explanation guided by abductive methods and inference to the best explanation. This description also includes a new re-interpretation of true scores in classical test theory afforded by the author’s measure-theoretic mental test theory development—for a particular test-taker, the variation in observed test-taker scores includes measurement error and variation attributable to the different ecological testing settings, which aligns with the explanation-focused view wherein item and test performance are the object of explanatory analyses. The final main section of the essay describes several methodological innovations in explanation-focused validity that are in response to the tensions and changes in assessment in the last 25 years.

### TABLE OF CONTENTS

1. INTRODUCTION.....	3
1.1. The Zeitgeist of the Late 20th to the Early 21st Century in Assessment Research.....	3
1.2. Purposes of the Paper .....	4
1.3. Structure of the Essay .....	5
2. EVOLVING DEFINITIONS OR DESCRIPTIONS OF VALIDITY .....	6
2.1. The Phrase “Validity Theory” Will Be Used Broadly and Inclusively .....	7
2.2. Distinguishing Validity Theory and Validation Methods .....	7
2.3. Developmental Periods and Changing Definitions/Descriptions of Validity - Eleven	

\*CONTACT: Bruno D. Zumbo ✉ [bruno.zumbo@ubc.ca](mailto:bruno.zumbo@ubc.ca) 📍 University of British Columbia, Measurement, Evaluation, & Research Methodology Program, Department of Educational and Counselling Psychology, and Special Education, Vancouver, BC Canada V6T 1Z4

Definitions or Descriptions of What is Meant by the Term Validity .....	8
3. QUESTIONS OF HISTORICAL CHANGES AND PROGRESS SINCE EARLY 1900 ..	20
3.1. Philosophy of Scientific Realism as It Relates to Theory Change and Progress .....	20
3.2. Are There Distinct Periods of Development in the Concept of Validity and Validation Methods From 1900 to the Present? .....	22
3.3. Are There Observable Patterns and Trends in the Historical Record? .....	23
3.4. Have We Made Progress in Our Description or Definition of Test Validity? .....	26
3.5. Notwithstanding That No Single Definition of Validity Theory Emerged, Several of Them Reflect Explanation-Centered Views .....	31
4. SETTING THE STAGE FOR MY EXPLANATION-FOCUSED VALIDITY .....	37
4.1. What Motivated the Development of My Explanation-Focused View? .....	37
4.2. Context, Ecology, Diversity, and the Many Ways of Being Human.....	40
4.3. Recognizing and Quantifying Uncertainties in Test Validation and Assessment Research Practice .....	42
4.4. Initially, Classical Test Theory Seems Simple, but Its Description and Interpretation Have Changed Over Time and Is Now Aligned with the Explanation-Focused View .....	43
4.5. Some Remarks on Measure-Theoretic Test Theory .....	46
4.6. The Re-interpretation of the True Score of CTT is an Affordance of Measure-Theoretic Test Theory That is Important to My Explanation-Focused Validity and Assessment Research.....	48
4.7. How Perspectival Realism and Pragmatic Undercurrents of Conditionalized Realism Inform My Explanation-Focused Validity Theory and Assessment Research.....	57
5. DESCRIPTION OF MY EXPLANATION-FOCUSED VALIDITY .....	59
5.1. Explanatory Considerations in Test Validation and Assessment Research .....	59
5.2. Basic Ideas Underlying My Explanation-Focused Validity: Bridging the Inferential Gap, Abductive Methods, Inference to the Best Explanation, and Explanatory Coherence.....	59
5.3. Exploratory Factor Analysis, Latent Variable Regression Models, and the Pratt Index for Variable-Ordering as Examples of Explanation-Focused Validation Methods .....	61
5.4. The Ecological Model of Item Responding and Subtest or Test Performance: A Conceptual Model.....	62
5.5. An Ecologically Informed, In Vivo View Describes the Enabling Conditions for the Abductive Explanation .....	64
5.6. Test Validity in The Context of Concomitant Changes In The Value-Free Ideal in The Philosophy Of Science.....	65
5.7. Explicit Synthesis of Explanation-Focused and Argument-Based Approaches to Test Validation .....	68
6. METHODOLOGICAL INNOVATIONS IN EXPLANATION-FOCUSED VALIDITY ..	70
6.1. Third Generation DIF is About More than Just Screening for Problematic Items.....	70
6.2. An Entrée for Embracing the Many Ways of Being Human in an Explanation-Focused Framework.....	73
6.3. The Importance of, and Multiple Ways to Think About, Loevinger’s Two Test Validation Settings .....	75
6.4. Response Processes Are Important to Test Validation: Insights from a Broadened View .....	76
6.5. Test Validation as Jazz .....	78
6.6. Test-Taker-Centered Assessment and Testing and Test Validation as Social Practice: The Case of Inclusive Educational Assessment, Neurodiversity and Disability.....	79
7. CONCLUSIONS .....	80
REFERENCES.....	84

## 1. INTRODUCTION

In this paper, I reflect on test validity's past and future in light of this journal volume's theme, *Lessons from the Past, Visions for the Future*. The global rise of assessments since the late 20th century coincided with a period of rapid development and increased availability of computational sophistication. Even recently, we have seen openly accessible conversational AI systems, software for advanced statistical and psychometric analyses, Web 3.0 or the metaverse, and digital innovations in test delivery. Additionally, assessment design, delivery, and test validity have changed significantly from 1960 to now, along with social, political, economic, cultural, scientific, and technological changes that have shaped our world. As such, this certainly feels like an appropriate time for an “over-the-shoulder look” back at some key moments in assessment. It is advisable, if not illuminating, to set a course forward by at least glancing at where we have been, so this paper takes a retrospective look at assessment while looking forward to the horizon for a glimpse of what lies in store.

These tectonic shifts also changed test validity in educational and psychological measurement. After describing historical trends in the definition of test validity, I glance back mainly from an explanation-focused perspective (e.g., Zumbo, 2005, 2007a, 2009). For other perspectives on test validity history, see Hubley and Zumbo (1996) for a historical description focused on Messick's contributions, Jonson and Plake (1998) for a historical comparison of validity standards, Sireci (2009) for a historical analysis focusing on the *Standards for Educational and Psychological Testing* (referred to as the *Standards* henceforth; American Educational Research Association [AERA], APA, & National Council on Measurement in Education [NCME], 2014), as well as the six previous editions, and Kane (2001) for a brief historical review of construct validity with an emphasis on argumentation. Even a cursory glance at the corpus of the major books in our field and the contributions on the pages of the *International Journal of Assessment Tools in Education* or other scholarly research journals like it, such as *Educational and Psychological Measurement* or *Journal of Educational Measurement*, shows tremendous developments in validity theory, validation practices, assessment methodology, and applications since the 1960s. To be more concrete, I will analyze test validity in the context of the intellectual and commercial forces that shape assessment applications and developments in test validity and assessment research.

### 1.1. The Zeitgeist of the Late 20th to the Early 21st Century in Assessment Research

General historical practice does not define these terms precisely; however, “late 20th century” generally refers to the last quarter or third of the 20th century, whereas the “early 21st century” is the first two decades of the 21st century.

The late 20th and early 21st century saw a global increase in the use of assessments, tests, and instruments for various purposes in the social sciences based on educational and psychological measurement developments. In education, large-scale testing, longitudinal testing, individual assessment, and surveys coincided with a growing economy of global assessment and testing. Of course, it would be disingenuous to portray vigorous activity and busyness on its own as reflecting a rosy picture of assessment practices: The rapid changes in assessment theory and practice of the late 20th and early 21st century left some important issues unresolved or in the background. Reflecting on these changes in the assessment field, Zumbo (2019) draws these issues to the foreground in his description of the tensions, intersectionality, and what is on the horizon for assessments in education. Two strands of contemporary international large-scale education assessments often sit in tension.

On the one hand, developers and purveyors of such assessments and surveys, those employed and profiting from the testing and assessment industrial complex, desire to ensure that their assessment tools and delivery systems are grounded in our most successful psychometric and

---

statistical theories. They aim to do social good while serving their economic and financial imperatives. There is nothing necessarily untoward or ignoble in this goal; what Zumbo (2019) describes is just a social and economic phenomenon reflecting financial globalization and international competitiveness.

On the other hand, there is the increasing desire of those of us outside of the test and assessment industrial complex, per se, to ensure that the philosophical, economic, sociological, and international comparative commitments in assessment research are grounded in a critical analysis that flushes out potential invalidities and intended and unintended personal and social consequences. These two strands are not necessarily disjoint and are connected by a common body and goal.

With the tension described above in mind, this essay is written with the continued belief that this tension is important and healthy as it unites both strands in working toward a common goal of increasing the quality of life of our citizens globally.

## 1.2. Purposes of the Paper

As Zumbo and Chan (2014a) show via a large-scale meta-synthesis of the genre of reporting test validity studies across many disciplines in the social, behavioral, and allied health sciences, this research is largely uncritical in presenting their subject matter, rarely indicating what of many possible validation frameworks were chosen nor why (Shear & Zumbo, 2014). As hidden invalidities may undermine test score claims, this research should focus on the concept, method, and validation process since invalid measures may harm test takers.

The first purpose is to summarize major trends in how prominent validity theories conceptualize test validity from the early 1900s to the early 2000s. There are two general aims associated with this first purpose. The first aim is to provide some organizing principles that allow one to catalog and then contrast the various implicit or explicit definitions of validity. I look at those trends mainly from an explanation-focused perspective (Zumbo, 2005, 2007a, 2009). The second aim of the historical analysis is to examine the extent to which the major trends and changes in prominent conceptions of validity and validity theories in the assessment field targeted exposing and documenting possible hidden invalidities. I ask the important question: have the descriptions and definitions of validity progressed to a single definitive theoretical account since the early 1900s? Along the way, I aim to shine a light on the context of the intellectual and commercial forces that shaped the changes in test design, development, and delivery and the changes in validity theory.

The outcome of the descriptive and historical analysis of changes in test validity serves as the basis for the second purpose: describing my explanation-focused test validity and what I see on the horizon regarding methodological innovations emerging from the vantage point of my explanation-focused view of assessment research and test validity (Zumbo, 2005, 2007a, 2009) embedded within an ecological model of item responding and test performance (Zumbo et al., 2015), placing a centrality to test consequences and values, and what I refer to as the many ways of being human (Zumbo, 2018a). For this second purpose, I also revisit the earliest articulations of my explanation-focused validity (Zumbo, 2005, 2007a, 2009) to describe what I have not done hereto and situate those contributions within my developments in the mathematical models of test theory that shaped my views of test validity. I will also briefly describe philosophical and psychological ideas that shaped my thinking. This process results in what may be described as field notes that reflect the ideas, impressions, thoughts, criticisms, and unanswered questions as I continue to develop my explanation-focused theory of validity and accompanying statistical methods. Drawing a thread from what led up to the first description of the explanation-focused view in my *Messick Award Lecture* (Zumbo, 2005) and reflecting on my field notes allows for a fuller description of what I see on the horizon of

assessment research and test validity from the vantage point of my explanation-focused view.

Notably, the first two purposes are motivated by possible hidden invalidities that may undermine test score inferences and claims while focusing on the concept, method, and validation process since invalid measures may harm test takers. These two purposes of this essay draw to the foreground what Zumbo (2019) describes as the tensions, intersectionality, and what is on the horizon for assessments in education and psychology.

The third purpose reflects a broader goal to create space where test validity research and assessment research more broadly can be considered setting the disciplinary silos aside to create greater space for multidisciplinary in inquiries of assessment research, test validity, and validation practices. Like others before it (Zumbo, 2007a; Zumbo & Chan, 2014a; Zumbo & Hubley, 2017), this paper aims to be a countervailing force against the widespread phenomenon of assessment researchers creating what I refer to as *measurement silos* and fragmented knowledge. These measurement silos may obstruct knowledge-sharing across fields and hinder innovation. Working against these silos does not mean that field-specific assessment research is invaluable; quite to the contrary. Nevertheless, some assessment research should aim to speak across the measurement silos to enhance our understanding of measurement, reduce fragmentation among researchers by removing boundaries, and combine expertise from various fields to solve complex problems. In line with the broader objective, it is important to note that the terms assessment, test, measure, and instrument will be used interchangeably and in their broadest senses to mean any coding or summarization of an observed phenomenon.

Therefore, lest we fall into traditional camps and comfortable silos, validity applies equally to instruments used in large-scale educational examinations, tests for certification and licensure, psychological instruments, psychosocial education research, and the learning sciences, to name a few. Of course, this statement about the broad implications of this commonality is not meant to suggest that there are no unique features; instead, it shines a light on the fact that we have far more in common to learn from each other than the comfortable disciplinary silos may suggest.

### **1.3. Structure of the Essay**

Although this essay is not comprehensive, it aspires to be self-contained to provide the reader with the context of discovery and the motivating factors for developing certain validity theories and methods. The topics were selected to motivate the reader to embrace the challenges of contemporary assessment research and test validation described in the earlier sections.

This paper is organized into seven sections to meet its purposes. Section two describes the difference between validity theory and validation and describes the evolution of the definition or description of the concept of validity since the early 1900s. I investigate the development of the definition or description of the term “validity” as it relates to validity theory or test validation because, with few exceptions, what is offered in the historical record does not resemble a theory, per se, even in the most liberal understanding of what is a theory. Doing so allows me to cast a wide net as I investigate how validity theory has evolved since the early 1900s. Section three addresses the natural questions that arise from the over-the-shoulder look back at the history of validity: what are the changes, whether they reflect progress in our understanding of test validity, and, if so, what kind of progress is it? An explanation-focused view of test validation and validation methods emerges from the historical analysis, setting the stage for my explanation-focused view. Therefore, section four sets the stage by describing the necessary conceptual and psychometric preliminaries for a detailed description of my explanation-focused view of test validity. Section five describes the current version of my explanation-focused view of assessment research and test validity, the confluence of ideas that influenced its initial development, and how it has developed into a coherent research framework for test validity and assessment research. Section six describes what is on the horizon regarding



---

innovations in methodology supporting the explanation-focused. Section seven is the conclusion, in which I provide a brief reflection on issues discussed in the article.

## 2. EVOLVING DEFINITIONS OR DESCRIPTIONS OF VALIDITY

This section describes some key moments in the history of validity theory reflecting the changes in the conceptualization or definition of validity from the early 1900s to date. In the latter part of this section, I continue the theme of key moments in the validity history mainly from the lens of an explanation-focused perspective (e.g., Zumbo, 2005, 2007a, 2009).

It is advisable, if not illuminating, to set a course forward by glancing at where we have been. Drawing on historical and contemporary research in test validity, I argue that contrasting concepts of validity are important for understanding the sources, methods, and the variety of knowledge claims that emerge from them. The description of the historical trends will aid in exploring the general principles and challenges of validity theory and validation practices in education research and large-scale assessment rather than focusing on a specific domain such as science assessment or context such as international comparative surveys such as those administered by the OECD.

The question addressed in this section is: What is meant by “validity” in educational and psychological measurement by investigating how validity theory has evolved since the early 1900s? Of course, the reader must be mindful that for most of this essay section, I focus on the various descriptions and definitions of the term “validity” in test validity; however, more generally, I attend to validity theory. In many cases, no explicit definition is offered. Still, a definition of sorts is, in essence, implied through the description of what the authors mean by the concept of validity offered in various influential publications that other researchers have cited since the early 1900s.

To be inclusive and cast a wide net of the historical record, I investigate the change in (a) what authors present as definitions of validity or test validity, (b) descriptions of the term “validity” rather than definitions as they relate to validity theory or test, and validation, and (c) theories of test validity offered. However, it is notable that, with few exceptions, what is offered in the historical record does not resemble a theory, per se, even in the most liberal understanding of a theory. Zumbo (2009) found that what is described as “validity theory” in articles in research journals, book chapters, or textbooks is a *mélange* of the three options listed above, with the most common being descriptions of the term “validity.”

Given the vast array of approaches to test validity that have emerged since the early 1900s, Zumbo (2007a) provides an important cautionary note.

Integrating and summarizing such a vast domain as validity invites, often rather facile, criticism. Nevertheless, if someone does not attempt to identify similarities among apparently different psychometric, methodological, and philosophic views and synthesize the results of various theoretical and statistical frameworks, we would probably find ourselves overwhelmed by a mass of independent models and investigations with little hope of communicating with anyone who does not happen to be specializing on “our” problem, techniques, or framework. Hence, in the interest of avoiding the monotony of the latter state of affairs, even thoroughly committed measurement specialists must welcome occasional attempts to compare, contrast, and wrest the kernels of truth from disparate validity positions. However, while we are welcoming such attempts, we must also guard against oversimplifications and confusion, and it is in the interest of the latter responsibility that I write to the more general aim. (Zumbo, 2007a, pp. 71-72).

As Zumbo (2007a) remarked, reading the vast literature on validity theory and practice dating back to the early 20th century leaves one with the impression that the history of test validity and validation practices exhibits a pattern characteristic of a maturing science. One is left with the impression that the history of test validity reveals a growing understanding and a series of

unending debates on topics of enduring interest. An example of growing understanding is a change in language from (a) distinct types of validity to (b) types of validity evidence. This change from types of validity to types of validity evidence may seem a subtle semantic move. However, as described below, these implications substantially affect test validity and validation practices. In terms of unending debates on topics of enduring interest, an obvious example is whether consequences should play any role in test validity and validation practices.

### **2.1. The Phrase “Validity Theory” Will Be Used Broadly and Inclusively**

As we transition to section two of this essay, it is important to describe how I use the phrase “validity theory” throughout this essay. There are no single elements explicitly designated as being “validity theory” because the terms “validity” and “theory” are used quite broadly both in assessment and testing practice and in meta-level discussions about the measurement theory and test validity.

To avoid confusion, the phrase “validity theory” will be used throughout this essay, following its conventional use in the educational and psychological measurement field. I will follow suit if something is referred to as a validity theory in the research literature and textbooks.

In addition, for our purposes herein, whether it is a theory is less important than what is meant by term validity. Therefore, to avoid dwelling on whether something described as validity theory in the educational and psychological measurement literature and textbooks is a theory per se, the historical analysis in section two of this essay focused on defining or describing the conception of the term “validity” in the phrases “validity theory” or “test validity.” Depending on the kind or amount of description or definition of validity provided in the research literature, the focus is on the denotation, connotation, or both of the word or expression for validity.

In summary, for section two of this essay, I will follow suit and include it for analysis if the approach, perspective, or view of validity is described as a theory of validity in the educational and psychological measurement literature or textbooks. Likewise, it need not be described as a theory, per se, to be included in section two. This broad use of the phrase validity theory will allow me to be inclusive in meeting our objectives of the historical analysis reported in section two and subsequent analysis in section three.

### **2.2. Distinguishing Validity Theory and Validation Methods**

This backward glance at the development of the concept of validity, as it pertains to test validity, will be just that: a glance—our primary goal is to describe theories and methods for validation. Zumbo (2007a, 2009) reminds us that it is important to distinguish between validity and validation at the outset. In assessment, testing, and measurement, *validity* is properly understood as denoting the property or relationship we are trying to judge; *validation* is an activity geared toward understanding and making that judgment (Borsboom et al., 2004; Zumbo, 2007a, 2009). Zumbo (2009) and Shear and Zumbo (2014) remind us of the importance that a guiding rationale (i.e., validity) must play in selecting and applying appropriate analyses (i.e., validation), while Zumbo et al. (2023) highlight how failing to distinguish between validity and validation can lead to conceptual and methodological confusion.

Zumbo and Chan (2014a) documented that test validation studies reported in the published educational and psychological research literature rarely explicitly define (or describe) what they mean by validity for the purpose of their research. However, it appeared that the language tended towards discussing the validity of scores and inferences. Reporting test validity evidence without clearly defining validity in published validation studies tends to confuse validity theory and validation methods, as validity theory literature shows (e.g., Messick 1989; Shear and Zumbo 2014; Zumbo 1998, 2007a, 2009). Therefore, test validity and validation must be distinguished to prevent overemphasizing data analysis methods without a conceptual basis. To

make this less abstract, I will provide two examples. For instance, the multi-trait multimethod (MTMM) approach from Campbell and Fiske (1959) is a validation method that follows Cronbach and Meehl's (1955) construct validity theory, which the survey research literature does not always acknowledge. Likewise, as shown in Zumbo et al. (2023), the validation methods of cognitive interviews or think-aloud methods are loosely founded on the notion of validity involving an explanation for the item responses and a description of the response process. To be clear, in this latter example, as Zumbo et al. note, this theory of validity involves providing an explanation for the variation in responses to survey questions or test items. The validation method is the cognitive interview or think-aloud interview.

Not surprisingly, the systematic reviews of the genre of reporting test validation studies in education and psychological research by Zumbo and Chan found that validation practices' statistical and psychometric complexity has increased over time. However, key sources of validity evidence remain hidden or under-represented. In addition, the theoretical concepts of validity, such as those reflected in the *Standards* and the framework described by Kane (2006, 2013) or Messick (1989), do not guide the validation process.

As Shear and Zumbo (2014) highlight, the systematic review of the genre of reporting practices for validation studies in research journals in their chapter, and overall in Zumbo and Chan (2014a), suggests two important implications in practice.

- First, as Messick (1995) warned, two primary threats to the validity of score interpretations are construct underrepresentation and construct irrelevant variance. For instance, a systematic study of test validity evidence based on response processes used by test takers (Zumbo et al., 2023) or the consequences of test interpretation and use (Hubley and Zumbo, 2011) could provide key evidence needed to shine a light on these currently mostly hidden threats to validity.
- Second, without a clear guiding theory of validity, it is hard to judge if a validity research program has met its goals. The absence of a guiding theory of validity also makes it difficult to compare findings from different validity studies that may have different aims. It undermines the *Standards'* statement that validity is “the most fundamental consideration in developing and evaluating tests” (AERA et al., 1999, p. 9) because the meaning of validity may be unclear. Different validity concepts can guide validation research, such as those reviewed above. However, more clarity is still needed on specific validation methods that can assess test scores according to these validity concepts.

In summary, to better understand the interplay between validity and validation, in this essay's subsequent sub-sections, we explore the various definitions or descriptions of validity offered in the research literature since the early 1900s and the validation methods implied by each definition. As we transition to the description of the developmental periods and changes in the definitions or descriptions of the concept of test validity, it bears repeating that I take a strong position here and elsewhere (Shear & Zumbo, 2014; Zumbo, 2009; Zumbo et al., 2023;) that one needs to describe what they mean by “validity” to go hand-in-hand with the methods used in the process of validation. I believe that my position is warranted because, by and large, test validation studies reported in research journals do not report being guided by any theoretical orientation, validity perspectives, or validity theory (Zumbo & Chan, 2014a, 2014b). Most troublingly, the extensive body of theoretical research literature on test validity, described below, or the *Standards*, are rarely mentioned or cited in the over 700 published test validation studies in research journals examined in Zumbo and Chan (2014a).

### **2.3. Developmental Periods and Changing Definitions/Descriptions of Validity - Eleven Definitions or Descriptions of What is Meant by the Term Validity**

The following eleven definitions or descriptions of the concept of validity- what the term “validity” means and how it is used- trace the historical development of educational and



psychological measurement. The documentation of the explication of the locution “validity” in educational and psychological measurement since the early 1900s and comparing it within a historical context shows how these continue to evolve and inform contemporary validation practices.

To avoid misunderstanding, before introducing the eleven definitions or descriptions of what the term validity means, it is important to note that I do not mean “definition” to mean scientific or operational variants thereof. In addition, I do not consider it a type of essentialism for definitions, nor does it involve a commitment that the assigned meaning agrees with prior uses (if any) of the particular description or definition of validity. Although these ways to consider the definition or description of validity may be interesting and may even provide insights, they would take me away from the more general purpose of this essay. Instead, my brief description of the evolution of the definitions or descriptions of “validity” in test validity in educational and psychological measurement is guided by ideas in speech-act theory, particularly what Searle (1969, 1979) describes as propositional acts that are clear and express a specific definable point, as opposed to mere utterance acts, which may be unintelligible sounds and illocutionary acts that tell people how things are.

In this essay, I expanded upon my project *tracing the evolution of the prominent conceptions of validity from the past century* with the intent of investigating the evolving conceptions of test validity’s impact on contemporary validity theory and validation practices (Shear & Zumbo, 2014; Zumbo & Padilla, 2020; Zumbo & Shear, 2011; Zumbo, 2010). Rather than approaching the task of tracing the descriptions and definitions of the concept of “validity” in test validity naively of linguistic theory, descriptions of speech-acts and a method described in Searle (1979) guided me. That is, I followed speech-act theory loosely, using it as a general framework rather than a strict rule.

The method I use in this essay is, in a sense, empirical. I studied and documented the language used in published articles, book chapters, and books in prominent conceptions of validity dating back to the early 1900s. I also documented the types of illocutionary points explicating the locution “validity.”

What follows in the next subsection of this essay builds on Shear and Zumbo (2014), which lists historical periods for concepts of validity and corresponding validation methods.

### ***2.3.1. A test is valid if it measures what it is supposed to***

The origins of this description of validity are typically described as the early 1900s. However, it is notable that there was no description of validity, per se, during this period; rather, the concept of validity is implied in the description of what makes a test valid.

The validity description during this period is embodied in Buckingham's (1921) and Curtis's (1921) descriptions of a test as valid if it measures what it is supposed to. Curtis writes: “[t]wo of the most important types of problems in measurement are those connected with the determination of what a test measures, and of how consistently it measures. The first should be called the problem of validity, the second, the problem of reliability” (p. 80). Similarly, Buckingham writes in the context of intelligence tests: “By validity I mean the extent to which they measure what they purport to measure. If for educational purposes we define intelligence as the ability to learn, the validity of an intelligence test is the extent to which it measures ability to learn” (p. 274).

Three points are noteworthy; first, these descriptions suggest that validity is a property of a test rather than a test score or inference. Second, these definitions of validity entail no single implied process or method of test validation. However, Curtis and Buckingham suggest considering the test scores’ associations with other variables as possible statistical information informing the judgment of validity without indicating how and what that statistical information may

provide the researcher. Third, remarkably, these first two points are enduring—see a definition of validity offered in the early 2000s by Borsboom et al. (2004, 2009).

### ***2.3.2. Validity is about establishing whether a test is a good predictive device or short-hand for a behavior***

During the two decades between the world wars (1918 to 1939), behaviorism was North American psychology's dominant school of thought. Influenced by early behaviorists (e.g., Hull, 1935; Watson, 1913), the dominant view of psychology during this period was partly a response to earlier forms of introspective methods and psychoanalysis embracing a science of human behavior. Behaviorists criticized both introspection and psychoanalysis for being subjective, unscientific, and unreliable. Behaviorists of this period argued that psychology should focus only on observable and measurable behavior and not on mental processes that could not be directly verified, rejecting that innate factors, such as instincts or drives, determined behavior.

This form of behavioral psychology, claiming that psychology is the science of human behavior, significantly impacted education and educational and psychological testing and measurement. Most notably, test scores were mostly considered signs or predictive devices for some future or alternative behavior. Validity is about establishing whether a test is a good predictive device or short-hand (criterion validity). Shear and Zumbo (2014) quote Angoff (1988, p. 20), who writes: “Consistent with other writers at that time, Bingham defined validity in purely operational terms, as simply the correlation of scores on a test with “some other objective measure of that which the test is used to measure (Bingham 1937, p. 214)”. Importantly, operationalism and operational definitions are invoked in this concept of validity.

This concept of validity suggests a specific, although limited, method of validation, which is the correlation of test results with a criterion. These criteria assessments frequently tend to forecast future actions or results, such as success in the workplace or college. In short, the received view of validity during this period is about establishing whether a test is a good predictive device or short-hand (criterion validity); therefore, a test is a predictive device or a shorthand. Regarding validation methods, one establishes whether a test is a good predictive device or short-hand. Therefore, the primary validation evidence is criterion correlation and prediction.

### ***2.3.3. The proliferation of “Types” of validity***

Huble and Zumbo (1996) describe the period between the 1930s and the late 1960s in test validity as intellectually vibrant, with many creative and innovative developments. This scholarly era in educational and psychological measurement was marked by encouraging various views to flourish and debate and being immersed within a central motivation for the activity.

In the 1940s and 1950s, many social and behavioral scientists felt the need and demand to have their field recognized as a science. However, a science demands that "things" (more specifically, behavior, affect, or cognition) be measured, and with measurement, one needs to have validity. Thus, many of the changes seen in the area of validity have come from work in psychological measurement that was motivated by this movement. (p. 210)

It is important to note that newer concepts of validity do not replace earlier ones in evolving the concepts of validity. So, by this period, the earlier views that (i) the test is valid if it measures what it is supposed to, and (ii) that validity is about establishing whether a test is a good predictive device or short-hand for behavior are still present and vibrant. Therefore, the criterion-based validity approach held its grip on test validation until the mid-1900s – and, not surprisingly, it reappears regularly throughout the history of validity and even presently.

This view is perhaps best reflected in Anastasi's (1950) characterization of the concept of

validity: "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" (Anastasi, 1950, p. 67). For example, if a test is designed to measure intelligence, the criterion could be academic achievement or occupational success. She stated that any claim that a test measures something beyond its criterion, such as an abstract construct or trait, is speculative and not based on empirical evidence. She argued that a test can only be validated by comparing it to a measure of the behavior or outcome the test intends to predict or explain, a specific criterion; Anastasi also pointed out that both the test and the criterion are samples of behavior, and many variables, such as motivation, mood, or situational factors, may influence either or both of them. Therefore, she suggested that test scores should be operationally defined in terms of empirically demonstrated behavior relationships rather than theoretical concepts.

Anastasi made a compelling case for a narrow description of test validity relative to the criterion or prediction on which it is based. However, for various reasons, dissenting views began to emerge that the criterion view was insufficient to capture the various uses and settings in which tests were being used. So, from the 1930s to the late 1960s, we see a proliferation of many types of validity. Sireci (2020) provides a rich snapshot of the different validity terms used in the seven AERA, APA, and NCME *Standards* versions- described as "categories" or "types" of validity in the 1952 and 1954 versions.

For instance, some psychological phenomena are abstract and do not have such a criterion or prediction. This instance shows the cracks in a restrictive adherence to behaviorism alone but also includes personality and clinical aspects that may affect the test scores. For example, in contrast to the narrow view of a test criterion, Guilford (1946) makes the case that "[i]n a very general sense, a test is valid for anything with which it correlates" (p. 429). I interpret this more expansive view to mean that a test potentially has as many validities as there are (significant) correlations.

As another indicator of the unrest and dissatisfaction with the narrow criterion definition during the 1930s to the late 1960s, Rulon (1946), Cureton (1951), and Lennon (1956) made a case for, defined, and extended the idea of content validity. Rulon argued that some tests (such as certain educational tests) are obviously valid because, by design, an inherent property allows them to be taken at face value. Rulon provides an example of tests that have this inherent or intrinsic validity (which are obviously valid) as educational tests "... in which the material presented to the student is the kind of material which constitutes the objectives of instruction, and in which the operation required of the student by the test situation is the operation which the school is trying to train the student to perform on such material" (p. 295). See Sireci (1998) for a thorough description of content validity development that continues to reflect the key concepts and issues.

Hublely and Zumbo (1996, p. 209) aim to capture the essence of this period. They described validity during this period as having many different types of validity available, and one chooses the type or types of validity most relevant or most easily obtainable to validate one test or assessment. This strategy of selecting one of several types of validity evidence can be seen in the best light as opportunistic and providing prima facie evidence, which in this setting does not mean that it proves or establishes validity but rather a fairly weak but essential claim in the early stages of a validation plan. Alternatively, selecting one of several types of validity evidence can be in a much worse light as somewhat haphazard (Zumbo & Chan, 2014b, p. 322).

Looking back at Hublely and Zumbo's description of having many different types of validity available, and one chooses the type or types of validity most relevant or most easily obtainable from today's perspective, it is apparent that perhaps without intending to, Millman (1979) reflected the emergent view of validity from the latter part of the 1960s onward: "... in judging any test, it is the use or interpretation of the scores that determines the appropriate indicators of test validity and reliability. Method follows function" (p. 75). As Hublely and Zumbo note, in

this statement, Millman appears to represent what many test developers seem to believe: Only certain types of validity, in the parlance of the time, need to be shown for different purposes.

#### **2.3.4. Cronbach and Meehl's 1955 description of construct validity**

Two interrelated key changes are reflected in the advent of Cronbach and Meehl's (1955) highly influential paper. First, if earlier in the 1900s, educational and psychological tests and assessments were considered predictive devices or shortcuts (or short-hand) for a behavior, then the period surrounding Cronbach and Meehl's contribution to test validity, a dominant view came to flourish that these tests and assessment were considered a structured way of "visualizing the unseen" through the self-report of test-takers. Reflecting a second related central change, as Shear and Zumbo (2014) note, researchers in the early history of validity wrestled with ways to determine "if a test measures what it is supposed to," as we noted, test scores also came to be seen increasingly in a behavioral light. Validity and validation in the first half of the twentieth century are often described as primarily empirical and possibly even atheoretical (Angoff, 1988).

Importantly, I wish to be careful not to assert that any criteria or observations can be theory-free. So, although I do not accept that any judgment or procedure of this nature can be completely atheoretical, I accept that these judgments and procedures would have reflected assessment theories such as projective or empirical criterion-keyed approaches (Hubley & Zumbo, 2013) that were hotly contested at the time. Likewise, the claim of being "atheoretical" could also refer to competing psychological theories; in particular, the early stages of what we would call a cognitive revolution began to replace psychoanalysis and behaviorism as the dominant approaches to studying psychology. In this sense, one could interpret Angoff's characterization of being "atheoretical" less controversially, that the intent was to take a neutral position concerning the competing psychological theories of the time.

Finally, although I do not accept that any judgment or procedure of this nature can be completely atheoretical, I accept that these judgments and procedures were based on what, upon reflection, Cronbach (1988) described as a weak program of construct validity I described above wherein any correlation of the test score with another variable is welcomed as validity evidence that also gave rise to the increasing array of "types" of validity and was driven primarily by the validation methods used rather than by a theoretical framework of validity. Partly, in response to this, The Technical Recommendations for Psychological Tests and Diagnostic Techniques (APA, 1954) introduced four aspects of validity: content validity, predictive validity, concurrent validity, and construct validity.

The American Psychological Association Committee on Psychological Tests found it necessary in the early 1950s to consider broadening the then-current definition of validity to accommodate the interpretations assigned to assessment in personality, abnormal, and clinical psychology. As Cronbach (1989) notes, a subcommittee of two members, Paul Meehl and Robert Challman, was asked to identify the kinds of evidence needed to justify the "psychological interpretation that was the stock-in-trade of counselors and clinicians" (p. 148). Cronbach goes on to state that Meehl and Challman introduced the notion and terminology of construct validity, which was incorporated in the 1954 Technical Recommendations (American Psychological Association, 1954). The concept of construct validity was more fully described by Cronbach and Meehl (1955).

The purpose of their influential article (Cronbach & Meehl, 1955) was to explain their concept of construct validity. As Shear and Zumbo state, although initially introduced along with content, criterion-related predictive, and criterion-related concurrent as a fourth "type" of validity, construct validity also brought a shift in perspective. Construct validity was initially intended to guide evaluating test score interpretations when no adequate criterion or content

definition was available. Using the philosophical and scientific principles of logical empiricism (Zumbo 2010), Cronbach and Meehl (1955) outlined an approach to articulating and testing a proposed nomological network, of which test scores were one observable result. Given that Cronbach and Meehl variously refer to both “construct validity” (p. 281) and “construct validation” (p. 299), their description of construct validity is not easily distinguished as either a definition of validity or a process of validation. For example, Cronbach and Meehl clearly articulated how one might gather evidence during the validation process. However, they also emphasized that “Construct validity is not to be identified solely by particular investigative procedures, but by the orientation of the investigator” (Cronbach and Meehl 1955, p. 282).

It should be noted that since its introduction in the field, many authors refer to construct validity as the most important characteristic of a test, but it is seldom defined. A clear statement of what a construct is and the logic of construct validation was presented by Cronbach and Meehl (1955). These authors wrote:

A construct is some postulated attribute of people, assumed to be reflected in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct. We expect a person at any time to possess or not possess a qualitative attribute (amnesia) or structure, or to possess some degree of a quantitative attribute (cheerfulness). . . . Persons who possess this attribute will, in situation X, act in manner Y (with a stated probability). The logic of construct validation is invoked whether the construct is highly systematized or loose, used in ramified theory or in a few simple propositions, used in absolute propositions or probability statements. We seek to specify how one is to defend a proposed interpretation of a test . . . .” (p. 247)

In short, a measure is valid for a construct when it produces results that can be interpreted regarding the construct definition under consideration.

Reflecting on the widespread and nearly immediate uptake of construct validity, Zumbo (2021, 2023a, 2023b) stated that some confusion arose among assessment practitioners and researchers from the fact that tests that are construct-valid provide information about (i) the study participant in terms of the construct and (ii) how the construct definition itself can be strengthened or extended. For some, the latter is counterintuitive: How can a previously constructed valid test provide information about strengthening or extending the construct definition? Distinguishing these two types of information and recognizing the importance of the second type is notable for two reasons. First, it is consistent with a key point made by the philosopher of science van Fraassen (2008, 2012), who highlighted in his study of the history and philosophy of measurement that the theory of the phenomenon and its measurement cannot be answered independently of each other, and they co-evolve. Second, this co-evolution is an important, yet largely unspoken, feature in my theory of validity and validation as an integrative cognitive judgment involving a form of contextualized and pragmatic best explanation that the practice of test validation will (should) inform the construct, competency, or attribute we posit to be measuring. This theme will be picked up again later in this essay.

Importantly, for the primary purpose of this essay to draw attention to an explanation-focused view of validity, Cronbach and Meehl state that the problem faced by assessment researchers is “What constructs accounts for variance in test performance” (p. 282); “Determining what psychological constructs account for test performance is desirable for almost any test” (p. 282); “A numerical statement of the degree of construct validity would be a statement of the proportion of the test score variance that is attributable to the construct variable” (p. 289).

As noted by Cronbach (1971), since the advent of construct validity, researchers in education and psychology have generally leaned toward the nomological network conception of psychological terms. It is argued that a construct is admissible if properly anchored in a nomological network. Thus, many pieces of evidence must be used to support a claim made



from a score from a test or assessment. Cronbach and Meehl's (1955) introduction of construct validity could reasonably be interpreted as the first nudge in the scientific direction to developers of psychological assessment and assessment researchers by providing an alternative to the prevailing operationalist and criterion-based approaches to test validity. In practice, however, shortly after its introduction, construct validity came to be viewed as a more abstract and global form of validity, even though it was meant to move in the opposite direction towards a deeper understanding of dispositions and the trait concept of that period that were poorly theorized in the psychological assessment of the time- as evidenced by the need to establish the American Psychological Association Committee on Psychological Tests as described above.

Importantly, as suggested by Cronbach (1988), a strong program was presented as the ideal. Along with this came an emphasis that validity and validation were about evaluating proposed interpretations of test scores rather than the test itself, a fundamental tenet of modern validity theory (Sireci, 2009; Zumbo, 2007a). As Shear and Zumbo note, despite this call for a holistic framework of scientific inquiry, validity remained a fragmented concept, and the type of validity one demonstrated was most often a product of the method used to document validity (Hubley and Zumbo 1996).

Kane (2001, pp. 321 – 326) provides a clear description of the setting of construct validity theory and a rich analysis of its strengths and weaknesses. Among Kane's insights that are important for the current essay is that there was a lack of clear criteria for the adequacy of validation efforts. Likewise, he states:

The basic principle of construct validity calling for the consideration of alternative interpretations offers one possible source of guidance in designing validity studies and in restraining empirical opportunism, but like many validation guidelines, this principle has been honored more in the breach than in the observance. (p. 326)

To be fair, Cronbach and Meehl (1955) did not aim to clear the field and describe a single view of validity (that would come later); their paper did not do much to slow down the proliferation of types of validity. As Hubley and Zumbo (1996) describe, in 1966, the validity terms predictive and concurrent were subsumed and replaced with criterion-related validity (Angoff, 1988). Thus, a trinitarian concept of validity emerged, as described by Hubley and Zumbo.

Although the trinitarian concept of validity prevailed historically, other types of validity have been proposed. Indeed, during the 1940s and 1950s there was a proliferation of different conceptions and delineations of validity. Some of the other validity types proposed include Guilford's (1946) factorial and practical validity, Mosier's (1947) face validity, Gulliksen's (1950a) intrinsic validity, and Anastasi's (1954) proposal of face, content, factorial, and empirical validity. (p. 210)

While the trinitarian concept of validity initially aided in elucidating validation procedures, it has, over time, produced unfavorable consequences for testing practices. It oversimplifies and crudely groups various data-gathering procedures meant to contribute to understanding what a test measures. Although there is some disagreement about whether the trinitarian concept was meant to introduce three aspects of validity (Guion, 1980) or three types of validity (Angoff, 1988), the three came to be viewed as separate entities. Guion (1980, p. 386) described these "as something of a Holy Trinity representing three different roads to psychometric salvation," meaning that at least one type of validity is needed. However, one has three chances to get it, a take-home message that continued unabated from the last period described above.

### ***2.3.5. Loevinger clears the way forward to construct validity as the whole of validity***

Into the 1960s and 1970s, even after the highly influential theoretical articulation of construct validity by Cronbach and Meehl (1955), anyone wishing to conduct test validation research would find themselves overwhelmed by a mass of independent concepts of validity and "types" of validity practices and investigations with little hope of communicating with anyone who does

not happen to be specializing in “our” problem, techniques, or framework.

With clarity of intellectual purpose and clear writing, Loevinger’s (1957) proclamation "since predictive, concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view" (p. 636) figuratively wrangled the proliferation of concepts and methods resulting from the “wild west” spirit of the period. Thus, our evolving definition of validity changed when Loevinger’s (1957) “construct of validity is the whole of validity” gained more popular support in the 1970s and the work of individuals such as Messick (1975), who argued that to properly judge the appropriateness, meaningfulness, and usefulness of an inference or claim based on a test score; one must have evidence of what the test score means or represents.

Loevinger (1957) makes the following points that are, for the most part, largely ignored in the validity theory research literature.

Thus, in place of the classification of validity proposed in the Technical Recommendations, it is here recommended that two basic contexts for defining validity be recognized, administrative and scientific. There are essentially two kinds of administrative validity, content and predictive-concurrent. There is only one kind of validity which exhibits the property of transposability or invariance under changes in administrative setting which is the touchstone of scientific usefulness: that is construct validity[*sic*]. (Loevinger, 1957, p. 641)

Neither the Technical Recommendations nor Cronbach and Meehl gave a formal definition of construct validity. In the former paper the term was introduced as follows: "Construct validity is evaluated by investigating what psychological qualities a test measures, i.e., by demonstrating that certain explanatory constructs account to some degree for performance on the test... Essentially, in studies of construct validity we are validating the theory underlying the test" (121, p. 14). (Loevinger, 1957, p. 641)

Cronbach and Meehl's introduction of the term was: "Construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not “operationally defined.” The problem faced by the investigator is, 'What constructs account for variance in test performance?' (20, p. 282) (Loevinger, 1957, pp. 641-642)

These distinctions and concepts will play a more central role as validity evolves. With the publication of a crucial article by Cronbach and Meehl in 1955, the construct model, which strongly focuses on construct validity, was introduced and moved toward in the early 1950s. Similarly, Loevinger (1957) made the crucial point that every test, if for no other reason than the fact that it is a test and not a criteria performance, underrepresents its construct to some extent and contains sources of irrelevant variance. The focus on observable behavior, theories of learning, and psychology's relatively recent split from psychoanalytic and introspective methods are reflected in the early- to mid-1900s in the validity history. The early stages of what we now refer to as the cognitive revolution of the 1970s were evident in the 1960s.

The period post-Cronbach and Meehl, mostly the 1970s to the present, saw the construct validity model take root and saw the measurement community, led by efforts of Sam Messick, delve into a moral and consequential foundation for validity and testing by expanding to include the consequences of test use and interpretation.

### ***2.3.6. Messick’s influence on test validity until the turn of the twenty-first century***

Discussing test validity and assessment research from the mid-1970s until the twenty-first century is challenging without considering Sam Messick’s views at length. His impact looms so large on this topic that most discussions of validity between 1975 and 2000, in some senses, are extensions, responses to Messick’s earlier writings, or both. Most certainly, my explanation-focused view embracing the many ways of being human that emerged in the late 1990s, described in a later section of this essay, is a case in point.

Messick (1975, 1980, 1988, 1989, 1995, 1998, 2000) articulated a unified view of validity in

---

several publications. He was clear that validity is about the inferences, interpretations, actions, or decisions based on a test score, not the test itself. It refers to the degree to which accumulated evidence supports the intended interpretation of test scores for the proposed purpose. Moreover, validity is about whether the inference one makes is appropriate, meaningful, and useful given the individual or sample with which one is dealing and the context in which the test user and individual/sample are working. That is, one cannot separate validity from the sample from which or the context in which the information was obtained (Zumbo, 2009).

Messick (1972) makes an early case for the importance of psychological processes, which he later called substantive validity evidence, in a paper largely ignored in the test validity literature. He states that one of the main challenges for psychology is to translate psychological theories from words to rules, making clear the structure of thought and behavior. Creating sequential models of psychological processes is essential, and factor analysis can reveal their key components. Factor analysis finds a few variables from consistent individual differences in complex behaviors, showing their relationships. Factor analysis also validates traits and provides the functional method to validate laws. This multivariate experimental method is tested from the literature and in connection to the nature and formation of psychological traits and complex processes in learning, problem-solving, and creativity. He showed that evidence for the role of factors of cognition and personality in influencing those complex performances has been increasing, forming a foundation for the final step of detailed model building.

Messick provided the most extensive consideration of consequences in assessment and testing. In the following extended quotation, Hubley and Zumbo (2011) highlight several critical points about Messick's unified view of validity relevant to considering social consequences.

Under the unified view, validity is all about the construct and meaning of scores. The validation process involves presenting evidence and a compelling argument to support the intended inference and show that alternative or competing inferences are not more viable. One refers to types of validity evidence rather than distinct types of validity. Furthermore, evidence is intended to inform an overall judgment; therefore, validation is not meant to be just a piecemeal activity. Messick and others (e.g., Hubley & Zumbo, 1996; Zumbo, 2007a, 2009) have strenuously argued that validity cannot rely solely on any one of these complementary forms of evidence in isolation from the others.

Finally, validation is an ongoing process. The unified model provides us with a regulative ideal that gives us something to strive for and governs our validation practice (Zumbo, 2009). However, as Messick (1989) points out, "Because evidence is always incomplete, validation is essentially a matter of making the most reasonable case to guide both current use of the test and current research to advance understanding of what test scores mean" (p. 13). Thus, we can think of this process as similar to repairing a ship while at sea (Zumbo, 2009).

The consequences of testing refer to the unanticipated or unintended consequences of legitimate test interpretation and use (Messick, 1998). There are two aspects to the consequential basis of testing: value implications and social consequences. Some writers have argued that social consequences have no place in validity; their argument tends to be based on a misconception that social consequences are about test use and, in particular, test misuse. First, the focus is on consequences, not use. Second, Messick (1998) did not view test misuse or illegitimate test use as part of the consequences of testing. Indeed, although they might be important concerns, he saw the consequences of test misuse as irrelevant to the nomological network and score meaning and thus outside of construct validity and the validation process.

As I will describe in more detail later in this essay, the aspect of Messick's theorizing that perhaps most reflects his thinking is the consequential basis for interpretation and use. Nevertheless, it is often misunderstood (Hubley & Zumbo, 2011). The consequential basis is

not about poor test practice. Instead, the consequences of testing refer to the unanticipated or unintended consequences of legitimate test interpretation and use (Messick, 1998).

Social consequences of legitimate test use can be positive or negative, and both are important in terms of validity. While the test developer and test user are often more concerned about unanticipated negative or adverse effects resulting from test use, Hubley and Zumbo (2011) argued that one must consider positive effects when considering validity and score meaning. Again, from a validity standpoint, the focus is on effects traceable to sources of invalidity, such as construct underrepresentation and construct-irrelevant variance. Because these consequences contribute to the soundness of score meaning, they are an integral part of construct validity and the validation process (Messick, 1989; 2000).

In summary, as Shear and Zumbo (2014) state, in an attempt to bring together these various strands of validity and validation that still dominated discourse about validity theory into the early 1970s, Messick (1989) provided the following definition of validity: “Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (p. 13). As Zumbo and Shear state: “While this definition of validity does not entail a single approach to validation, three widely accepted guiding tenets are that (a) numerous sources of evidence can contribute to a judgment of validity, (b) validity is a matter of degree rather than all or none and, (c) one validates particular uses and interpretations of test scores, rather than a test itself.” (p. 95)

### ***2.3.7. Embretson’s construct validity is a universal and interactive system of evidence, emphasizing construct representation and nomothetic span***

Embretson (1983, 2007) described construct validity as a universal and interactive system of evidence, emphasizing construct representation and nomothetic span. Embretson’s framework is the first of the descriptions of validity that I encountered that explicitly implies a research method to investigate the claims made in the framework. This feature of Embretson’s framework is a strength because it supports the interpretation of formal cognitive modeling and correlational techniques, among others.

In addition to this institutionalized definition of validity presented by the AERA, APA, and NCME (1999) *Standards*, Zumbo (2010) highlights that the research program by Embretson (1983) could be read as a response (or follow-up) to Cronbach and Meehl (1955). She characterizes her view of validity as a “universal and interactive system” (Embretson, 2007, p. 452). Much like Loevinger before her, it appears that Embretson aimed to bring clarity of purpose to construct validation described by Cronbach and Meehl.

What has come to be called response processes evidence in support of validity is a central aspect of Embretson’s conception of validity (Zumbo & Hubley, 2017). As noted by Hubley and Zumbo (2017), Embretson generously gives the nod to Messick’s early (1972) claim that there is a need in the psychometric field to develop models of psychological processes that underlie test performance (Whitely, 1977). Embretson (1983) proposed that construct validity is comprised of two aspects: (a) construct representation and (b) nomothetic span. Construct representation involves identifying theoretical mechanisms (e.g., processes, strategies, knowledge stores, metacomponents) that underlie test items or task performance. In contrast, nomothetic span involves relationships between the test score(s) and other variables. In the parlance of the *Standards* (AERA et al., 1999, 2014), one might think of construct representation as falling under the response processes’ source of evidence and nomothetic span as falling under the relations to other variables’ source of evidence. As Hubley and Zumbo note, Embretson (1983) saw construct representation as concerned with test scores’ meaning. In contrast, the nomothetic span has to do with the significance of test scores. Furthermore, she



and her colleagues argued that the theoretical mechanisms can be examined using task decomposition methods from information processing (Embretson et al., 1986).

Embretson's conception of validity draws heavily on the notion of construct representation versus nomothetic span; the former deals largely with cognitive processes and modeling, and the latter with observed relationships (Embretson, 1983, 1998, 2007). This framework provides substantial emphasis on modeling cognitive processes and internal test characteristics while also providing a framework for integrating multiple forms of evidence. Zumbo et al. (2023) show Embretson's influence among the earliest descriptions of response processes as validity evidence in the transition from the behaviorist to information processing and early traditions of cognitive psychology. As Zumbo et al. state, these early signs of information processing research led to a nascent kind of cognitive-psychometric modeling of response processes initiated in the mid-1970s by Susan Embretson (Whitely) (e.g., Embretson, 1983, 1984, 1993; Embretson et al., 1986; Whitely, 1977).

### ***2.3.8. Haig's and Zumbo's explanation-focused views of validity***

Haig (1999) argued for adopting a broad explanationist outlook on construct validation in which the generation, development, and different forms of abductive reasoning carry out a comparative appraisal of theories. They make a sound case that validation is a form of abduction and that the process of discovery (for example, see Thagard, 1992) shows that scientists often reason from empirical generalizations to explanatory theories to infer and evaluate possible explanations in an abductive way. Haig (in press) provides a full and rich articulation of his explanation-centered view of validation, which historically should be read as the long-awaited response to Cronbach and Meehl's (1955) articulated from a contemporary philosophy of science. A central theme in Haig's (in press) recent views is the important turn away from nomological networks to pragmatic theories and their evaluation by explanatory means.

Zumbo (2005, 2007a, 2009) independently introduced explanation-focused views of test validity in which construct validity centrally involves making inferences of an explanatory nature, highlighting inference to the best explanation (IBE). This reliance on explanation and IBE was presented contra the dominant mode of construct validation framed as hypothetico-deductive empirical tests in line with Cronbach and Meehl and those scholars who advocated that view. The view of validity described in a later section of this essay that is meant to guide our assessment research reflects Zumbo's perspective on construct validity: "[e]xplanation acts as a regulative ideal; validity is the explanation for the test score variation, and validation is the process of developing and testing the explanation" (2009, p. 69).

As described earlier in this essay, Zumbo's explanation-focused view is central to the purpose of this essay; therefore, it will be more fully articulated in the third section of this essay.

### ***2.3.9. Two clear departures from the modern, unified approach to validity***

Two clear departures from the contemporary unified approach to test validity have drawn attention and advances since 2000. As described by the authors when these views were introduced, these two views reflected bold strategies aimed to strip down the more elaborate notions of validity reflected largely by developments from Cronbach and Meehl to Messick and reflected in the Test Standards.

Lissitz and Samuelsen (2007) describe validity as related solely to internal test characteristics. They write: "Together, we suggest that these essentially internal characteristics (reliability and content validity) be called the internal validity of the test, and all other characteristics be considered essentially external matters" (p. 446). They aimed to outline a concept of validity with more clearly developed and practical validation methods. Their conception is well-suited to modern methods of content validation, cognitive modeling, and reliability analysis (p. 445). While they recognize the importance of additional sources of evidence, they seem to consider



these distinct from a determination of validity.

Borsboom et al. (2004, 2009) proposed a radically different definition of validity, which, in short, aims to extract construct validity from the theories of validity. They state their point clearly: "... a test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure" (Borsboom et al. 2004, p. 1061). Importantly, the contemporary view of validity in the tradition of a unified view per Cronbach-Meehl-Messick describes validity as a property of test scores or inferences, not as suggested by Borsboom et al. that validity is a property of tests. Borsboom et al. offer validating tests by stating formal cognitive theories, developing tests from these theories, and empirically investigating response behavior.

### **2.3.10. Schaffner's construct progressivity assessment**

Schaffner (2020) introduces an approach to test validity that applies construct validity. Still, for reasons he develops in his article related to his conceptualization of the concepts of "truth" and "validity," it is better thought of as construct progressivity assessment (CPA). Schaffner (2020) proposed that construct validation is a process of epistemic appraisal of competing models or theories, assessing various models or theories using empirical and extra-empirical standards that speak to a model's theoretical virtues.

For this essay, Schaffner's view of "construct validity" is not only a recent offering in the long line of construct validity approaches in educational and psychological measurement but also an important reminder of the distinction between two ideas that are often presented as intermixed in contemporary test validation practices: (a) the validation of constructs as theory appraisal, more generally, and (b) test validity. In addition, we are reminded of the contingency of validity claims. The clarity of Schaffner's exposition helps bring these two points to the forefront.

**2.3.10.1. Distinguishing the Validation of Constructs as Theory Appraisal and Test Validity.** Schaffner (2020) begins his article by describing a variation on the widely accepted description of constructs in the main educational and psychological measurement. Concepts like intelligence frequently refer to general, abstract, and putatively explanatory entities, and these types of entities are often generally termed constructs. He goes on to state that:

"... considerable investigatory efforts involve assessments of the reliability and validity of those constructs. Determining whether such constructs are valid—whether they are fictions and fantasies or are "real" (at least in the sense that they have appropriate explanatory power, utility, and strong evidential support)—can be approached from a variety of perspectives and traditions" (p. 1214).

A close read of Schaffner's description is that the constructs that we typically seek to validate, such as intelligence, must be validated indirectly. So, in the process of validation, we are looking for correlates of constructs, and the constructs put an interpretation on the observed behavior.

Nothing is inherently amiss with Schaffner's description; however, it highlights that his conceptualization of "validity" is focused on validating the construct. This conceptualization is not unreasonable; after all, it is quite reasonable to read "construct validity" as the process of the validation of constructs and, to some extent, the theories that contain them. Cronbach and Meehl, on the other hand, as I described earlier in this section of the essay, more narrowly define constructs as "some postulated attributes of people, assumed to be reflected in test performance" (Cronbach & Meehl, 1955, p. 247), as such, constructs are tied to what may be thought of as test validity. Unfortunately, Cronbach and Meehl and their interpreters are not always as clear in their distinction of (a) validating constructs more generally as theory evaluation and (b) test validity, which is more closely tied to the process depicted in the quotation early in this paragraph of Cronbach and Meehl. Of course, one could interpret

Schaffner's CPA as test validity akin to Cronbach and Meehl, where Schaffner's "observed" behavior is the item response. Schaffner makes this turn to test validity without fanfare when he relates CPA to Kane's argument-based approach.

Some validity theorists have worked to distinguish the validation of constructs from test validity. Borsboom et al. (2004, 2009) also highlight this point and suggest separating construct validity from measurement concerns. Haig (in press) provides a thoroughgoing and accessible discussion of concerns regarding the mixing of the evaluation of theory (construct validity as the evaluation of constructs as a kind of theory evaluation) and test validation (akin to Cronbach and Meehl's description in the quotation earlier in this section) and presents a reconciliation of this often unaddressed issue. Haig initially argues for the separation of the validation of constructs and test validity for strategic reasons, allowing him to highlight the importance of his preferred interpretation of construct validation as theory appraisal but, in the end, arrives at a fruitful reconciliation. In the end, Haig makes the case that construct validity and test validity should be brought back together by invoking developments in coherentist epistemology and a theory of explanatory coherence- see Haig (in press) for details. In closing, Haig also notes that Schaffner's wide view on theory assessment could reasonably encompass an explanation-focused view, particularly inference to the best explanation.

**2.3.10.2. Unlike Some Validity Theories That Imply A Universality of Validity Claims, CPA Is Temporally Contingent.** The second matter that Schaffner's description highlights is that, unlike some validity theories, for example, those who argue validity as a property of a test, CPA is temporally contingent. This temporal contingency recognizes that test validity may change depending on data from newer instruments and methodological advances (p. 1224) and, therefore, is not a universal claim. This contingency is also noted by Cronbach and Meehl (1955), Messick (1989), Hubleby and Zumbo (2011), and Zumbo (2007a, 2009), among others. This contingency is a central point of this essay: test validity must address how well the inferences, uses, or both of a test or assessment travel across time and place.

### 3. QUESTIONS OF HISTORICAL CHANGES AND PROGRESS SINCE EARLY 1900

The focus of this section of the essay is the analysis of the patterns of change and documenting major themes in the historical record of changes in validity theory reported in section two of this essay and whether these changes reflect progress in our understanding of test validity, and, if so, what kind of progress is it? By interrogating the assumptions and evidence behind the different conceptions of validity and validity theory and characterizing the diversity of scientific practices, we advance our understanding of how the notions of validity and validity theory work and decipher what kinds of answers they deliver. To my knowledge, no analysis of this kind has been reported in the research literature.

To better understand the interplay between validity and validation, section two of this essay describes the definitions or descriptions of the term validity offered in the test validity research literature since the early 1900s and the validation methods implied by each definition. Recall that these descriptions or definitions and their aligned validation methods characterize what is commonly referred to as a validity theory in educational and psychological measurement textbooks and research journals. This section analyzes the historical record of changes in the concept of validity and validity theory and validation since 1900.

#### 3.1. Philosophy of Scientific Realism as It Relates to Theory Change and Progress

##### 3.1.1. *Why should we be concerned with the philosophy of scientific realism?*

By the standard account in the philosophy of science, claims regarding realism, anti-realism, and nonrealism take center stage when one asks questions about theory change and scientific progress. However, this has largely been ignored in accounts of theory change in test validity, leaving the reader uncertain of how the author(s) ground their analysis and unable to interpret

or adjudicate the conclusions appropriately. This need to ground the research of theory change in a philosophy of scientific realism becomes particularly important if one aims to go beyond the most basic cataloging of concepts (e.g., Kuhn, 1996). To avoid this kind of uncertainty and confusion, I describe my stance on the philosophy of scientific realism as it relates to questions that arise during the analysis of changes in the descriptions and definitions of the concept of validity in assessment and testing and differences in validity theory since 1900. This description also allows me to describe how my stance on philosophic realism has shaped and informed my explanation-focused view of validity theory, validation, and assessment research practice in a later section of this essay.

A description of realism that captures its varieties in the philosophy of science literature is too complex to address in the present essay. However, in its simplest form, it is common to consider three dimensions of realism—a commitment to a mind-independent world, literal semantics, and epistemic access to unobservables. Philosophers of science have given much attention to the question, “What is scientific realism?” but have not agreed on a clear answer. There are many varieties of realism and various postpositivist antirealisms that challenge them.

I agree with Haig (2014, 2019) that it is fair to say that scientific realism, of some form, remains the dominant position in the current philosophy of science. I share the view of Kincaid (2000) and Haig and Evers (2016) that we cannot settle realism issues in the social sciences by philosophical arguments that judge whole domains of science; local formulations, not global arguments, can help us better understand realism in the social sciences like educational and psychological testing and assessment.

### ***3.1.2. Giere’s perspectival realism highly influences my views***

Adapting and paraphrasing Stathis Psillos’ (2022) opening remarks on theory change paints a vivid picture of our task in this essay section. In section two, we saw that descriptions and definitions of validity and validity theories seem to have an expiration date. A number of descriptions and theories that once were dominant and widely accepted are currently taught in the history of assessment and measurement, if at all. Will this be the fate of the current dominant approaches? Is there a pattern of radical theory change as the assessment and measurement science grows? Are validity theories abandoned en bloc? Or are there patterns of retention in theory-change? Are some parts of approaches to validity and validity theories more likely to survive than others? Moreover, what are the implications of all this for the scientific image of educational and psychological measurement and testing?

The image painted by Psillos evokes questions of scientific realism because it challenges the idea that science is a cumulative and progressive enterprise that converges to the truth. If scientific theories change radically over time and are incompatible with each other, how can we be confident that our current theories are true or approximately true? How can we explain the success of past theories that were later discarded or modified? How can we justify our inferences from observable phenomena to unobservable entities?

Many discussions of scientific progress, particularly outside of the philosophy of science literature, base their analysis of changes over time on the often unstated and undifferentiated realist idea that the advancement of science involves a build-up of truth about a common domain of entities. In our case of changes in the conceptualization of test validity and validation practices, this would include zeroing in on, getting closer and closer, to a single approximation to a true (correct) conception of validity. Although I continue to see some valid points in the constructivist critiques of realism, my view is highly influenced by Giere’s (2006) “perspectival realism.”

It is important to note that my views on the philosophy of scientific realism continue to reflect a substantial pragmatic component. Schaffner’s (1993) “conditionalized realism” shaped my

earliest theoretical developments in validity theory and continues to do so. However, my current leanings are closer to perspectival realism. Schaffner's conceptual clarity helped me navigate the choppy philosophy waters and currents wherein I do not embrace a strong anti-realist stance in my assessment research and theorizing. Still, I also reject a wholly committed (which I may describe as naïve) realism. As such, I resist the insistence of some forms of realism that perception provides unmediated access to the material world. In this way, I agree with Schaffner that we do not have any direct intuitive experience of the certitude of scientific hypotheses or theories. I continue to have an appreciation for several points raised by Nickles (2017), Fine (1984), and van Fraassen (1980, 1985) regarding the debates about realism in the philosophy of science and a growing appreciation for several central themes in Fine's description of a "natural ontological attitude."

Shaffner's pragmatic philosophic stance is on display as it motivates his argument (Schaffner, 2020, p. 1217) that it would be better to approach the arguments in Kane's (2013) approach to validation, which I describe later in this section of the essay, in the spirit of the American philosopher John Dewey's logic of inquiry (Dewey, 1938) than Toulmin's (1958) formulation of arguments in general, which Kane elaborated as part of his notion of an interpretation/use argument (IUA) analysis of construct validity. Schaffner states that this use of Dewey's logic of inquiry has the advantage of being closer to the kind of presentations we encounter in scientific review articles. As support for this Deweyian recommendation, Shaffner points out that the close relationship between Dewey's discussion of warrants and assertions and Kane's discussion of warrants and claims has already been observed in the test validity literature (Stone & Zumbo, 2016; Zumbo, 2009). Finally, we can take as a demonstration the nuanced implications of the philosophy of realism; Schaffner (2020, p. 1217) states that Toulmin's reference to truth differs from Dewey's theory of truth because, "for Dewey, there is no preliminary or even accessible truth, but only ongoing processes aimed at increasing the support of claims."

### **3.2. Are There Distinct Periods of Development in the Concept of Validity and Validation Methods From 1900 to the Present?**

Let us recall that the first purpose of this essay is to summarize major trends in how prominent validity theories conceptualize test validity from the early 1900s to the early 2000s to provide some organizing principles that allow one to catalog and then contrast the various implicit or explicit definitions or descriptions, denotations, and connotations of the concept of validity. In many assessment research and practice settings, these definitions and descriptions of the concept of validity in test validity travel under the umbrella of "theories of validity." Although there is not widespread agreement among philosophers of science about how to characterize the nature of scientific theories, the developments in Cronbach and Meehl (1955) may be the first that is likely to pass as a theory per se.

Unsurprisingly, educational and psychological measurement has largely inherited the spirit of a cumulative view of scientific progress that inspired epistemological views that regarded human knowledge as a process. Not only was the cumulative view of scientific progress an important ingredient in the optimism of measurement's roots in the positivist program of accumulating empirically certified truths, but science also promotes progress in society.

Similar to Shear and Zumbo (2014), I propose that we consider what appears to be four somewhat distinct periods of validity praxis and theorizing. The reader should remember two noteworthy points in my description of these four periods. First, I am not suggesting distinct historical periods and a natural linear step-wise progression toward our current thinking. I am not suggesting some evolution to the best theories. Second, I use the term *praxis* herein to convey a distinction between practice and theory, highlight the application or use of the knowledge and skills, and also reflect some of what is, in essence, the convention, habit, or

custom of validity work of the periods.

A brief description of the four periods of validity practice and theorizing follows.

1. The early- to mid-1900s were dominated by the criterion-based model of validity, with some focus on content-based validity models.
2. The mid-1900s to the late 1960s saw the introduction of, and move toward, the construct model, emphasizing construct validity, a seminal piece being Cronbach and Meehl (1955).
3. The period post-Cronbach and Meehl, mostly the 1960s to the end of the 1990s, saw the construct model take root and saw the measurement community delve into a moral foundation for validity and testing by expanding to include the consequences of test use and interpretation (Messick, 1975, 1980, 1988, 1989, 1995, 1998).
4. A period since about 2000 in which the debate about validity and validation has started up again after a quiet time post Cronbach's and Messick's programs of research.

Focusing more on the methods used for validation, a cluster of three periods may be created. From the early 1900s to the 1930s, the criterion view was the dominant method of test validation. The key element is validity as correlation or prediction of either an objective measure of that which the test is used to measure a criterion or anything for which it correlates. The mid-1930s to the late 1960s saw the proliferation of the multiple "types" of validity and the belief that we are validating the measures in the psychological and education research literature and the early versions of the APA/AERA/NCME Standards. As Hubley and Zumbo (1996) highlighted, the period from the 1960s to the end of the 1990s saw continued use of the language of *types of validity*, including, for example, discriminant validity, convergent validity, face validity, as well as the methodological developments beyond the simple validity coefficient (a correlation) to patterns among planned validation studies in the multi-trait multi-method matrix. Notably, the notion of constructs took root and construct validity as the accumulation of evidence had its dominance from the 1960s to the end of the 1990s but peaked in the mid-1970s and is still ongoing.

The landmark paper in this tradition is Cronbach and Meehl (1955), who described construct validity and the explicit use of the nomological network to establish the meaningfulness of a test or measure. The APA/AERA Standards (1974) reflect this dominant view of the time: construct validity is based on accumulating research results: formulate and test hypotheses using a hypothetico-deductive form of inferential reasoning. Cronbach's (1971) and later view of validation (and perhaps validity) as evaluation and, in some sense, a process of social, rhetorical arguments was a notable break in formalism and from his earlier collaboration with Meehl in 1955.

### **3.3. Are There Observable Patterns and Trends in the Historical Record?**

#### **3.3.1. Two patterns and a trend in the historical record**

Two patterns, defined as repeated occurrences of an event or behavior and a trend reflecting the general direction in which something is developing or changing over time, were discerned in the historical record in section two of this essay.

Notably, the two patterns are consistent with those reported in their historical analyses in Hubley and Zumbo (1996) and Zumbo et al. (2023). The first pattern is that the educational and psychological measurement literature continues to repeat the problematic practice of conflating a concept of validity with the validation method or process of validation. As Zumbo (2007a, 2009) notes, separating the concept of validity from the test validation process is important. For example, according to this view, validity, per se, is not established until one has an explanatory model of the variation in item responses, test scores, or sub-scale scores and the variables mediating, moderating, and otherwise affecting the response outcome, separating the concept of validity from the process of validation points to the fact that by focusing on the validation



---

process rather than the concept of validity we have somewhat lost our way as a discipline. This example is not meant to suggest that the activities of the validation process, such as correlations with a criterion or a convergent measure, dimensionality assessment, item response modeling, or differential item or test functioning, are irrelevant.

On the contrary, it points to the fact that the information from the validation process needs to be aligned with the concept of validity. The validation process must be directed toward supporting the concept of validity and not the end goal itself. I aim to re-focus our attention on why we are conducting all of these psychometric analyses: to support our claim of the validity of our inferences from a given measure.

For example, one continues to see the claims that validity is a correlation with a criterion or its more sophisticated-sounding kin that conflates the concept of validity with the estimation of variance components or component ratios using a cross-classified mixed effects model. Another example is described by Zumbo et al. (2023) for when substantial validity evidence from response processes is conflated with the method used to attain it: response processes are cognitive probes/think-aloud methods. In both instances, either no description or definition of validity is provided, and the conflation is obvious, or the description of validity provided lacks meaningful content beyond self-evident platitudes that do not advance our understanding of test validity. The second pattern is closely related to the second; finding an explicit definition of validity is uncommon. With a few exceptions (e.g., Borsboom et al., 2004; Cronbach & Meehl, 1955; Haig, 1999, in press; Zumbo, 2007a), the definition of validity being offered is, in essence, implied rather than stated. For this reason, I have referred to those views without explicit definitions as reflecting *descriptions and definitions* of validity or the concept of validity that arrived at through close study of the source material.

Finally, the trend that stands out is the tendency for a greatly expanded view of validity and validation practices over time. From the 1900s to date, the conceptions of validity became more expansive compared to the definitions in the first half of the 1900s, and so too have the entailed validation methods. As described in section two of this essay, during the 1940s and 1950s, there was a proliferation of different conceptions and types (or kinds) of validity. Indeed, one commonly encountered recommendation for test validity is that almost any information gathered in developing or using a test is relevant to its validity. Information deemed relevant was labeled another type of validity because it contributes to our understanding of what the test measures. For example, although many textbooks and theoreticians from the 1980s onward called for practitioners to stop using “face validity” because it was not considered validity, per se, there are recent examples in which it is of value as validity evidence (Galupo et al., 2018). Contributing to the expansive view of validity once introduced into the literature and they take root, validity types (or kinds) never become extinct because they may be of value in boutique cases of validation. Perhaps rightfully, validity theorists and validation specialists have become hoarders of validity types (or kinds) because “you never know when it will come in handy.”

The mid-1950s to the late 1990s witnessed many theoretical developments as the construct model introduced by Cronbach and Meehl (1955) took root, was modified, and expanded. It is worth repeating that amid this acceptance and development of expansive conceptions of validity theory and validation methods, and we saw two descriptions of test validity (Borsboom et al., 2004; Lissitz & Samuelsen, 2007) gain attention in the early 2000s that aimed to strip down the more elaborated notions of validity that had evolved and took root since the mid-1950s.

Thus, the dominant view of validity that emerged over the first 120 years was an increasingly expansive concept, moving from distinct “types” of validity that could be demonstrated through a single correlation coefficient to more nuanced theories that advocate that validity is no longer seen as a static property of tests but rather as an integrated judgment about the degree of the justifiability of inferences we make based on test scores (Messick, 1989). As validity became

increasingly expansive, it became more complex, giving rise to debates about what evidence is needed in different contexts. The late 1990s and the first few years of the 2000s marked a time of active development of validity theory and validation practices in educational and psychological measurement.

### ***3.3.2. Kane's argument-based approach in response to the complexity due to the greatly expanded view of validity and validation practices***

An influential development in validity theory in response to the complexity due to the greatly expanded view of validity and validation practice is the articulation of an argument-based approach to validation (Cronbach 1988; Kane 1992, 2006, 2013; Shepard 1993). Since the early 1990s, Michael Kane has been instrumental in fully developing and articulating an argument-based approach adopted in many large-scale testing programs. A key contribution of Kane's argument-based approach to validation is that it provides a disciplined and transparent methodology for establishing a validation plan, setting priorities, and interpreting validity evidence (e.g., Kane, 1992, 2001, 2004, 2006, 2013).

Notably, I do not include Kane's argument-based approach in the overview of validity concepts in this essay's second section because it does not derive from or require a particular definition of validity. Instead, it can be used as a methodology to support validation efforts guided by different definitions of validity. As Kane notes, the argument-based approach provides a "methodology or technology for validation" (Kane 2004, p. 136) rather than a definition of validity. As Shear and Zumbo (2014) note, Kane initially developed this method to support construct validity investigation, as Messick describes it (1989), and the 1999 Standards. It is consistent with those views of validity.

The argument-based approach grows from the notion that we validate inferences and uses rather than tests. We must clearly state the inferences and assumptions that move us from observed performances to proposed interpretations regarding a construct or its uses. In this light, Kane describes an interpretive argument, which clearly states the assumptions and inferences that move us from an observation to a final interpretation or decision. Then, in a separate process called a validity argument, we evaluate the plausibility of the proposed inferences and assumptions. Cronbach (1988), Kane (1992, 2006), Shepard (1993), and others advocate using an argument to frame or focus validation efforts and to clarify intended interpretations and uses.

I agree with Kane, who writes: "The main advantage of the argument-based approach to validation is the guidance it provides in allocating research effort and gauging progress in the validation effort" (Kane, 2006, p. 23). Some additional highlights of Kane's approach are different forms of interpretive arguments, the interpretive argument followed by the validity argument, and the distinction between descriptive and decision-based interpretations. Argument-based approaches have certainly embraced construct theories, but they foreground competencies.

As Zumbo and Shear (2011) note, we might compare the argument-based and explanation-focused approaches at a more conceptual level by posing the following question: Is an explanation an argument, or is an argument an explanation? There probably are multiple answers. If one approaches this question from informal logic (Sinnott-Armstrong & Fogelin, 2010), explanations are seen as types of arguments. There are at least two types of arguments: justificatory and explanatory. Distinguished largely by purpose or use rather than form, explanatory arguments provide an explanation of why or how something we agree about has happened; how did we arrive at a particular interpretation? Justificatory arguments provide reasons for belief; why should I accept the proposed interpretation? Focusing on the purpose of the argument brings our attention to who the audience is, which in some settings may be important. Returning to Kane's argument-based approach, one may consider the interpretive

argument explanatory and the validity argument justificatory.

These two sorts of arguments, justificatory and explanatory, often have similar forms, moving through chains of inferences. However, their purposes and the context in which we use them will often differ. There is an interesting parallel here between focusing on using a test to guide validation work; similarly, we can focus on using the argument to guide our construction of the argument.

Zumbo (2007a, 2009, 2017) notes that in terms of the process of validation (as opposed to validity itself), the statistical methods, as well as the psychological and more qualitative methods of psychometrics, work to establish and support the inference to the best explanation (IBE)– i.e., validity itself; so that validity is the explanation, whereas the process of validation involves the myriad methods of psychometrics to establish and support that explanation. Interestingly, it is notable that IBE essentially combines the justificatory and explanatory sorts of arguments; first, we formulate an explanation, then a justificatory argument to convince us it is indeed the best possible explanation.

Although it is clear how the validity argument serves to evaluate the pieces of the interpretive argument, what standards ought to be used to judge whether the interpretive argument, in context, is complete or serves its purpose (Messick, 1995)? Zumbo and Shear (2011) suggest that perhaps by conceptualizing the interpretive argument as explanatory, we gain a new set of criteria (for explanations) to evaluate our interpretive argument. By framing the two parts of the validity argument as explanatory/justificatory, we can leverage various frameworks for evaluating explanations in the service of developing our interpretive argument. In addition to Kane’s clarity, coherence, plausibility of inference, and assumptions, “[i]mplicit assumptions can be particularly harmful because they may be left unexamined” (Kane, 2006, p. 29).

Zumbo and Shear state that just as measures are fallible (hence the need for validation), our arguments are also fallible. Moreover, some arguments may be solid in one context but not another. Therefore, we need an analogous procedure to be sure our arguments are sufficient in a particular case, the same way we evaluate whether a test use or interpretation is sufficient in a particular context. Criteria for inference to the best explanations (think: selecting the best interpretive argument): “In sum, a hypothesis provides the best explanation when it is more explanatory, powerful, falsifiable, modest, simple, and conservative than any competing hypothesis” (Sinnott-Armstrong & Fogelin, 2010, p. 262).

### **3.4. Have We Made Progress in Our Description or Definition of Test Validity?**

The response to the question in the sub-section heading is not a straightforward “yes” or “no.” Although questions of this nature imply a binary response, the appropriate response in the case of the progress in test validity is: “Yes and no, it depends on the level of abstraction of the historical record.” Of course, the affirmative or negative responses need not be of equal force. The affirmative response will ultimately win the day in the question of progress in our description or definition of test validity, depending on the level of discourse that concerns the object itself, the concept of validity. I will briefly describe the subtle differences and variations that make it difficult to categorize in a straightforward response and unpack them below.

In short, the arguments regarding progress in test validity theory fall into two distinct levels of abstraction: the surface and the meta-level built upon it. Meta level is a distinction between levels of abstraction. The surface level, sometimes called the object level, is usually about a specific issue. At the same time, the meta-level is about general principles or “arguments about arguments.” At the surface level, one attends to particular failures to arrive at a single definition or description of the concept of validity as documented in the historical record in section two of this essay. That is, in support of the negative response to the question in the title of this subsection, 120 years of theoretical developments are marked by conceptual clutter that limits

the fields' cumulative progress. Furthermore, this conceptual clutter and lack of a singular definition of validity may result in choices among validity theories and validation methods determined by what is seen as fashionable trends. Although I continue to see some valid points, I do not find the details of these arguments at the surface level all that convincing.

The second level, a meta-level, provides clear evidence of progress toward a definitive statement about test validity that I derived from my analysis of the definitions and descriptions of validity from an explanation-based perspective (Zumbo, 2009). This second level also includes methodological considerations regarding the roles of the varieties of realism and anti-realism when making judgments of scientific practice.

### ***3.4.1. The surface-level analysis: Test validity theory has not progressed to a single definitive theoretical account***

It will be helpful to provide a few remarks about theory progression as a background to my analysis of the development of test validity since its earliest descriptions in 1900. Before the publication of Kuhn's highly influential book *The Structure of Scientific Revolutions* (1962, 1970), the widely held view that approaching psychological and educational research as science provided us with progress was viewed as development-by-accumulation of accepted facts and theories. Scientific progress was seen as accumulating new truths on top of the old ones, improving theories to match the truth, and occasionally correcting errors. This progress is guaranteed by the scientific method. As such, one should see progress toward a single definitive theoretical account of psychological and educational phenomena.

Although it is difficult to briefly summarize the complex and nuanced ideas offered in his books, not doing so would leave the reader missing an important part of the analysis of theory development. Kuhn's (1962, 1970) main idea is that science normally follows a "paradigm" that sets the problems and solutions for scientists. When a paradigm fails to solve some anomalies in the evidence or theory, science faces a crisis and may change to a new paradigm. This crisis and change is called a scientific revolution. Kuhn also argued that different paradigms are "incommensurable," meaning they cannot be compared or judged by a common standard. Incommensurability was one of the most contentious ideas in Kuhn's early work partly because it challenges some traditional views of scientific progress, such as the idea that later science builds on or gets closer to the truth than earlier science.

As described in section two of this essay, the evolution of test validity since the early 1900s has resulted in a plurality of definitions or descriptions of the concept of "validity" and the implied validation methods, therefore, a plurality of validity theories. At the surface level, there is no clear agreement on test validity. This surface-level analysis of the language and descriptions of test validity and validation practices provides ample evidence that progress has not drawn closer to a definitive statement about test validity, which suggests several possibly incommensurable validity theories. This lack of progress toward a definitive statement about test validity may be alarming to some assessment researchers influenced by Kuhn's (1970) developments because of a conviction they hold that multiple (possibly incompatible views of test validity) should not coexist, except during scientific revolutions.

Something is amiss when one compares the (surface-level) historical development of test validity since 1900 because there is no evidence of key positivist doctrines in the pre-Kuhnian (positivist) view of scientific progress. Likewise, if, for example, normal science progresses with a single view of test validity, there is no support for a Kuhnian view. One is left with the conclusion of the surface-level analysis that theories and activities of test validity and validation methods are pre-scientific or not scientific. Even if one accepts the claim that test validity is at a pre-scientific stage of development, in Kuhn's view, incommensurability can devastate the progress of validity theory and the practice of test validation. That is, in the third edition of *The*

---

*Structure of Scientific Revolutions*, Kuhn worked to clarify the concept of incommensurability, suggesting that, as applied to our context of the test validity, the plurality of incommensurable validity theories (a) undermines rational theory choice among validity theories, (b) leads to failures in communication, and (c) relegates rival validity theories and subsequent validity studies to different worlds (Kuhn, 1996, pp. 148-151).

Let us put some flesh on the bones of the incommensurability described in the previous paragraph to make it less abstract. When planning a validity study, many approaches to test validity are offered in the educational and psychological measurement literature. Choosing which one to use is like deciding what to wear for a night out on the town: it depends on the occasion, where you are going, your personal style, and what you want to communicate to others. With the metaphor of test validity “à la mode” in mind, we can imagine, for example, despite teased hair going out of style in the 1980s, a natural big hair trend is now à la mode and returning to fashion. In other words, the test validity equivalent to that sentence would be: Despite (defining validity as related only to item content) going out of style in the 1980s, a trend of (only reporting evidence related to content validity) is now à la mode and returning to fashion.

The metaphor of à la mode validity also has some face validity (forgive the pun) because a case could be made in the history of test validity that, in some cases, like the fashion industry, fashionable validity theories have been driven by the cult of personality of their designers and marketing campaigns. One wonders, for example, whether construct validity theory would have been taken up so quickly if it were not aligned with a major APA initiative and described by two eminent members of the psychological research community. Likewise, like the color of socks and scarves, there is no one true (correct) color choice.

In this vision of fashionable validity, à la mode, influential scholars, like designers and artists, use their talents and force of personality to advocate for a view of validity that appears de novo, responding to the particular demands or needs of testing scenarios such as projective tests of personality, clinical screening tests, or educational performance assessments. One could interpret Cronbach and Meehl as an instantiation of this precise motivation for a new test validity, construct validity.

The conclusion based on the surface-level analysis can be summarized as follows. The discipline of educational and psychological measurement has no visible singular strand of cumulative cognitive advances. At the surface level, validity theory is not just a multi-paradigmatic science. It is not limited to one single approach or perspective. Rather, it encompasses multiple paradigms, each with assumptions, methods, and criteria for evaluating validity. Therefore, at the surface level, validity theory is a complex and diverse field of inquiry requiring multiple lenses and perspectives to appreciate its richness and depth fully. As such, a plurality of definitions and descriptions of validity may be warranted given the many different purposes and uses of testing and assessment in varied settings involving potentially negative or positive immediate or short-term consequences, assessments or surveys designed for research purposes to large-scale assessment or testing programs, and ranging, for example, from relatively technologically advanced assessment programs to those that involve little technology. For example, the description of test validity offered in the early 1900s, that a test is valid if it measures what it is supposed to, can be found recently.

Most surely, even a cursory glance at section two of this essay leads the reader to conclude that the concept of validity has changed, as have the validation methods appropriate for those conceptions since the early 1900s. However, at the surface level, this change does not reflect a rejection of earlier concepts leading to a single approximation to a true (correct) conception of validity or validity theory. Against the background of changes in validity documented in section two of this essay, is there any reason to discuss scientific revolutions or counter-revolutions in



the historical analyses of concepts of test validity? Probably not, at least in the sense of Kuhn (1970, 1977). Kuhn challenged the common view of science as getting closer to the truth about nature by introducing new and controversial ideas, such as paradigms, scientific revolutions, and incommensurability. He described science as a problem-solving activity guided by paradigms, which are eventually replaced when they fail to deal with anomalies and a better paradigm emerges. However, whatever you may think progress looks like — an analogy between biological evolution and the evolution of science for expository reasons only, or epistemic iteration as a process by which knowledge claims are corrected or enriched — the surface level changes in the concept of validity from 1900 to date do not match these patterns.

In the following, I will summarize my outlook on the changes in the surface-level descriptions and interpretations of test validity. I take the view that there is not much prospect that the field of educational and psychological measurement will deliver a single, optimal surface-level description or definition of the concept of “validity” in test validity even in the next decade—the reason being that the last few decades of testing and assessment research has uncovered systemic complexity revealing hidden sources of invalidity, rather than a universal surface-level description or definition of the term “validity” in test validity.

### **3.4.2. The meta-level analysis: We have made important progress in test validity since the early 1900s**

It bears repeating that I do not find the details of these arguments at the surface level all that convincing. In this section of the essay, we will see that important progress in defining and describing validity theory and aligned validation methods has been made at the meta-level.

As Zumbo (2023b) states, there is an embarrassment of riches for test developers and users with more options or resources than one knows what to do when choosing among the test validation approaches and strategies. For each test, it is necessary to select the most appropriate method and, if necessary, modify it or create another method. Tailored for principled practices in test validation, Zumbo (2023b) states the following.

However, the embarrassment of riches does not mean we are in the wild west without rules and order. The Achilles heel of test validation is if the validation practices appear arbitrary, unjustified, capricious, and therefore vulnerable to missing hidden invalidity. Best practices are consequently defined in terms of choosing an approach and methodology that fosters transparency and justification for the choices one makes in the process of validation and an evidential trail that is both reproducible by test reviewers or other test developers, thus leading to the defensibility of the claims and uses/decisions made from the test scores. In short, the research journey is more important than the destination when judging best practices for test validation. (p. 103)

In summary, the changes in the description or definition of validity in test validity in educational and psychological measurement are best characterized by discontinuities and fashions that prevail over cumulative conceptual developments, constructive intellectual innovations, and repetitions. Nonetheless, these surface-level claims, although having some merit, are not convincing.

As shown by Zumbo and Chan (2014a), the reporting of validation studies in scientific journals has continued to grow unabated. Zumbo and Chan (2014c) documented the trend in the publication of validation studies between 1961 and 2010, with just over 300 publications between 1961 and 1965 and over 10,200 publications between 2006 and 2010. Certainly, some of that increase can be attributed to the rise in the sheer number of journals and researchers; however, the fact is that the field of measurement validity is growing in remarkable strides. Distinct approaches taken toward validation are difficult to discern in published research because, throughout most of the modern history of the field, researchers have presented research without explicit reference to a framework. At the same time, when considering what counts as validity evidence, Shear and Zumbo (2014) vigorously make the point that it is more important

---

that a validity theory be articulated and helps inform choices of validation practices than advocating that a particular concept of validity be adopted. Therefore, test validation practices can vary greatly, and there is no universal validation theory or method.

Two interesting questions arise when contrasting (a) the marked increase in the number of validation studies reported in research journals (Zumbo & Chan, 2014a) and (b) the negative view of progress in test validity since 1900 based on the surface-level analysis in the section above in this essay.

- Reflecting upon day-to-day contemporary test validation practices, what guides the decisions made during test validation studies' planning, conduct, and reporting?
- Moreover, what is one to make of the substantial number of validity studies and the amount of validity evidence reported?

It is important to note that the validity studies synthesized in the chapters of Zumbo and Chan are cited in substantive research to support new data collection with these tools. Substantive research claims are made (e.g., assessing the efficacy of interventions or programs) in education and psychology, so researchers find the test validation studies of value to inform later research using these instruments. Therefore, asking these two questions of validity theorists and assessment researchers would be interesting and valuable in investigating progress in test validity theory. In short, what do assessment researchers busily amassing an extensive body of test validation research literature know that test theorists do not?

Based on the over 700 validation studies included in our large systematic review of the genre of validation studies in research journals (Zumbo & Chan, 2014a), I would anticipate a difference of opinion and outlook between test validity theorists and practitioners. I anticipate that validity theorists' would tend to express the belief that test validity research works best when only one view of test validity allows assessment researchers to communicate easily and compare findings across other validation studies. In contrast, I would anticipate that the assessment researchers conducting and reporting validity studies on their tests and assessments of interest would express the belief that multiple views of test validity should coexist because they believe different types of validity are appropriate for different purposes and contexts of assessment. Assessment researchers may argue that no (single) universal definition of validity can apply to all tests and measurements. Instead, they may suggest that validity is a matter of degree and depends on the evidence and arguments supporting the test results' intended use and interpretation. They would also likely acknowledge that different views of validity may reflect different philosophical and theoretical perspectives on the nature of knowledge and reality. As such, from a practitioner's point of view, matters are not as pessimistic as the surface-level analysis of the change in validity theory may suggest, which contributes to why I do not find the details of these arguments at the surface level all that convincing.

The strongest evidence for why I do not find the details of the arguments of the surface-level analysis convincing is based on an investigation of meta-level progress in the definition and description of validity and aligned validation methods. I cannot stress enough that if we focus on progress since the 1900s, as we saw in the second section of this essay, there is undeniably great surface-level evidence supporting the lack of progress toward a single definition or description of the concept of validity.

As we transition to the meta-level analysis, a guiding question may be under what circumstances could we reasonably expect a single approach to or theory of test validity to suffice for a domain of educational and psychological phenomena like mathematics achievement or intelligence, respectively? An important step forward in addressing this question comes from reminding ourselves of the essential difference between surface and meta levels in comparing theories. The surface-level comparisons focus on the specific content of

different theories, expressed differently, on the observable and explicit “what” and “how” of each description or definition of the notion of validity or validity theory in section two of this essay. In contrast, meta-level comparisons focus on the underlying principles and frameworks that guide the different descriptions or definitions of the notion of validity or validity theory in section two theories. In its current use in this section of the essay, a principle or framework in the philosophy of science is a general guideline or criterion that helps evaluate the qualities and scope of scientific knowledge and methods. Many different principles and frameworks have been described, often reflecting diverse perspectives and assumptions about the nature and purpose of science: for example, empiricism, falsifiability, and parsimony or Occam’s razor.

At the meta-level, Zumbo’s (2009) initial theory comparison of Cronbach and Meehl (1955), Borsboom et al. (2004), and Zumbo (2007a), guided by the principle of scientific explanation, provides an argument that not only is theoretical progress possible but that there is preliminary evidence that it is, to some extent, already happening. I chose the principle of scientific explanation to guide the meta-level analysis because, as we saw in the historical record reported in section two of this essay, test validation has moved from a correlation or descriptive factor analysis to establish “factorial validity” as sufficient evidence for validity to an integrative approach to the process of validation involving the complex weighing of various bodies, sources, and bits of evidence, which naturally brings test validity and the validation process squarely into the domain of disciplined inquiry and science (Zumbo, 2007a, p. 72). Furthermore, in my view, seeking an explanation for our empirical findings is a hallmark of science.

A contemporary philosophical approach to science led me to a broad current view of scientific explanation and understanding (e.g., Friedman, 1974; Lipton, 2004; Persson & Ylikoski, 2007; Pitt, 1988; Salmon, 1990) encompassing many different kinds of scientific explanations rather than narrow views based on certain views of causation. A defining feature of the explanation-focused approach to theory comparison, described in this essay’s next section, is that it focuses on seeking explicit statements defining or describing the concept of validity or test validity and how one establishes it for each validity theory. The meta-level analysis reported herein aims to facilitate and motivate the further development of a science of assessment and testing development and research.

### **3.5. Notwithstanding That No Single Definition of Validity Theory Emerged, Several of Them Reflect Explanation-Centered Views**

There is no single agreed-upon definition of test validity; however, a group of eight approaches to test validity reflects an explanation-centered view of validity. Building on the case made in Zumbo (2009), the validity theories that focus on differing types of explanation and differing amounts of importance when describing their conceptualization of validity or validation include the following.

#### **3.5.1. Cronbach and Meehl**

Cronbach and Meehl (1955) described their notion of construct validity, which aims to provide an explanation for the test score variation using what they describe as a nomological network and invoking a variation on a covering law model of scientific explanation. One may interpret the concept of a nomological network as an interlocking system of laws that, in essence, constitute a theory. As such, constructs are like inductive summaries.

#### **3.5.2. Loevinger**

Loevinger’s (1957) scientific context of defining validity may reasonably be taken to focus on an explanation similar to Cronbach and Meehl’s. Notably, instead of being one type of validity amongst others, to Loevinger, construct validity was validity, that is, “... since predictive,

concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view” (Loevinger, 1957, p. 636).

### 3.5.3. *Messick*

Messick (1989, 1995, 2000) described his notion of substantive validity as one of six distinguishable aspects of his construct validity evidence, which Zumbo et al. (2023) describe as aimed at explaining the individual differences in the cognitive and behavioral processes involved in test performance.

### 3.5.4. *Embretson*

Embretson (1983, 1998, 2007) describes their notion of construct representation as largely dealing with cognitive processes and modeling related to response processes. Zumbo et al. (2023) describe Embretson’s validity theory as aimed at developing and testing explanatory cognitive-psychometric models of item response processes in support of test design and validation.

### 3.5.5. *Borsboom, Mellenbergh, and van Heerden*

In addition to Borsboom and his colleagues, Haig and Zumbo explicitly describe what “validity” means in their theories. This explicit description of “validity” greatly facilitates their presentation and comparison for this essay section.

Borsboom et al. (2004, 2009) rely on a causal model of explanation when they argue that a test is valid for measuring an attribute if, and only if, the attribute exists and variations in the attribute causally produce variations in the outcomes of the measurement procedure. They make a strong case that Cronbach and Meehl’s description of construct validity is problematic and should be abandoned to retain and strengthen the idea of test validity as the proper concern of validity and that it addresses (one may say, operationalizes) what they consider an important claim described in the early 1900s history of validity: a test is valid if it measures what it purports to measure.

A key idea in Borsboom et al.’s (2004) validity theory is their interpretation of the broad class of common factor models presupposes an underlying latent variable that gives rise to observed indicator variables, which may be item responses, ratings, or composite scores. The latent variable is then thought to correspond to some psychological attribute of interest – note that the authors describe why they avoid the word “construct” in their description. Although all we observe are its observed indicators, they assume that the underlying latent variable has causal efficacy. This key idea in Borsboom et al.’s theory of validity can be considered a literal interpretation of the path diagram of factor analysis where the arrows reflect actual causal paths. In short, Borsboom et al.’s validity theory considers the depiction of factor analysis in a path analysis as a theory of response processes. As I have observed (Zumbo, 2009), their definition of validity has virtue because it is, as the authors themselves acknowledge:

... a very tidy and simple idea that has a currency among researchers because it may well be implicit in the thinking of many practicing researchers. From my explanatory-focused view, relying on causality is natural and plausible and provides a clear distinction between understanding why a phenomenon occurs and merely knowing that it does—given that it is possible to know that a phenomenon occurs without knowing what caused it. Moreover, their view draws this distinction in a way that makes understanding the variation in observed item and test scores, and hence validity, unmysterious and objective. Validity is not some sort of super-knowledge of the phenomenon one wishes to measure, such as that embodied in the meta-theoretical views of Messick, Cronbach and Meehl, and myself, but simply more knowledge: knowledge of causes. (p. 73)

### **3.5.6. Haig**

Haig (1999, in press) argued for adopting a broad explanationist outlook on construct validation in which different forms of abductive reasoning carry out the generation, development, and comparative appraisal of theories. Key concepts in my interpretation of Haig's theory include (a) similar to Borsboom et al. distinguishing construct validity from test validity, where the former is thought of as an important form of test validity, (b) a shift in focus from construct validity to theory evaluation, (c) replacing the nomological network with a pragmatic view of theories, (d) abandoning the hypothetico-deductive method in favor of an explanation-centered view, and (e) appraising explanatory theories by employing the method of inference to the best explanation (e.g., Haig, 2019).

Although it was not presented as such, per se, I believe Haig (in press) is the strongest direct response to Cronbach and Meehl's (1955) construct validity in the educational and psychological research literature.

### **3.5.7. Zumbo**

Given that this theory of validity is the focus of the remaining sections of this essay, I will highlight three central features. First, as Zumbo (2007a) states, whereas validity is the property or relationship we are trying to judge, validation is an activity geared toward understanding and making that judgment. Zumbo argues on several occasions about the importance that a guiding rationale (i.e., validity) must play in selecting and applying appropriate analyses (i.e., validation) and that failing to distinguish between validity and validation can lead to conceptual and methodological confusion (Zumbo, 2007a, 2009; Zumbo et al., 2023). In doing so, they highlight the importance of having a clear concept of validity, which can guide the choice and use of validation methods.

Second, Zumbo's view of validity strongly emphasizes the centrality of explanatory inference. That is, validity is a matter of inference, and weighing evidence and explanatory considerations guides our inferences (Zumbo, 2007a). That is, as Zumbo (2009, p. 69) states, "Explanation acts as a regulative ideal; validity is the explanation for the test-score variation, and validation is the process of developing and testing the explanation." (2009, p. 69). Furthermore, invalidity distorts the meaning of test results for some groups of examinees in some contexts for some purposes, foreshadowing the view presented in Zumbo (2007b) and Zumbo et al. (2015) establishing the ecological model of item and test responding and for whom (and for whom not) the test or item score inferences are valid.

Starting with Zumbo (2007a), inference to the best explanation has played an important role in my explanation-focused view of test validity to generate and evaluate plausible explanations. The ecological model of item and test responding (Zumbo et al., 2015; Zumbo & Gelin, 2005) is central to establishing initial conditions, the facts or assumptions given at the start of abductive inference. They play an important role in determining the quality and plausibility of the abductive conclusion. Depending on the initial conditions, different explanations might be more or less likely, relevant, or consistent. Other abduction theories have different views on how initial conditions should be chosen, used, and updated in abductive inference. Some theories emphasize the role of background knowledge, prior probabilities, or explanatory criteria in selecting the initial conditions. Others focus on how new observations, feedback, or testing can revise or expand initial conditions.

Third, Zumbo (2007a, 2009) has described validity as a contextualized and pragmatic form of explanation. In this framework, validity is an emergent property that arises when an inference to the best explanation for observed test score variation supports proposed inferences and interpretations. Such a property depends upon the context of measurement and the context of interpretation and explanation. Thus, it centers on the role of values and consequences of



testing, including what I describe as the many ways of being human as it relates to assessment and testing.

### 3.5.8. Schaffner

Schaffner (2020) introduced the construct progressivity assessment (CPA) as a process of epistemic appraisal of competing models or theories, assessing various models or theories using empirical and extra-empirical standards that speak to a model's theoretical virtues. With an eye toward test validity, per se, the CPA approach may reasonably involve the appraisal of the competing explanatory models or theories of item or test score variation. Haig (in press) states that Schaffner's approach is a broad outlook on theory appraisal that may reasonably be taken to accommodate inference to the best explanation.

### 3.5.9. *Comparing the explanans and explanandum for the explanation-centered approaches*

In this section, I compare the explanation-focused validity theories regarding their explanations in terms of (a) what needs to be explained, the event to be explained, and (b) what contains the explanation, that is, the explanation of the event — as, for example, a cause, antecedent event, initial conditions, or necessary condition. The “explanandum” is the thing being explained, and the “explanans” is the explanation.

Of the eight validity theories that fit within an explanation-centered viewpoint, only a subset makes explicit and observable claims that allow me to ascertain the intended explanandum, explanans, or both. For example, Schaffner's CAP represents a broad view of theory appraisal; therefore, there is nothing amiss because the level of detail I am looking for is unnecessary and does not fit the purpose of Schaffner's (2020) paper.

I devoted attention to describing my definition of test validity because I hold as a first principle that if one wants to advance the theorizing and practice of measurement, I believe one needs to articulate what they mean by “validity” to go hand-in-hand with the validation process (Shear & Zumbo, 2014; Zumbo, 2007a, 2009). Where appropriate, however, I include Kane's (2006, 2013) argument-based approach to validation. However, as described earlier in this essay, by design, it does not incorporate a definition or description of validity. However, it is currently an influential view of test validation.

In my view of explanation, the relation between the explanandum and the explanans is considered from an abductive lens and an inference to the best explanation. In contrast, for Cronbach and Meehl (1955) and Borsboom et al. (2004), the relation is causal but reasonably taken to be deductive (a variant on the covering law) for the former and a causal claim of the sort described in the following for the latter.

What needs to be tested is not a theory about the relation between the attribute measured and other attributes but a theory of response behavior. Somewhere in the chain of events that occurs between item administration and item response, the measured attribute must play a causal role in determining what value the measurements outcomes will take; otherwise, the test cannot be valid for measuring the attribute. It is important to note that this implies that the problem of validity cannot be solved by psychometric techniques or models alone. On the contrary, it must be addressed by substantive theory. Validity is the one problem in testing that psychology cannot contract out to methodology. (p. 1062)

In the first sentence of this quotation, Borsboom et al. do away with Cronbach and Meehl's reliance on a nomological network very tidily and focus on the centrality of item and test response behavior. Borsboom et al. and my explanation-focused view of test validity have a commonality of purpose in the focus on response behavior. Still, beyond that, as described in this sub-section and the next three sections of this essay, the epistemological and ontological differences are substantial.

As it has impacted test validity, as Zumbo (2009) noted, there has been a long history of

competing ideas about what is and qualifies as an explanation in philosophy, with the deductive-nomological or covering law models garnering the greatest attention from the late 1940s to the late 1960s. As described earlier, Cronbach and Meehl (1955) rely on a variant of the covering law approach to explanation. As an alternative to covering law views, explanation has also been associated with causation more generally; an explanation is a description of the various causes of the phenomenon; hence, explaining is to give information about the causal history that led to the phenomenon. Borsboom et al. (2004, 2009) rely on a variant of this causal view explanation. In addition to covering laws and causal views of explanation, there is a third broadly defined view of explanation, often called the pragmatic approach, of which my explanation-focused view reflects a contextualized and pragmatic view of explanation; see Zumbo (2009) for a discussion of this view and its implications for test validation.

The basic idea underlying my explanatory approach is that understanding the item or task score variation would go a long way toward bridging the inferential gap between measurement scores and the constructs. One needs to know “what” they are measuring” and “what they are measuring along the way” because strict unidimensional “pure” unidimensional measures are highly unlikely in practice. This expectation is a tall hurdle indeed; however, as we saw earlier in this essay, the spirit of Cronbach and Meehl’s (1955) work was to require (causal) explanation in a strong form of construct validity.

I share with other validity theorists that validity is a matter of inference and the weighing of the evidence; however, in my view, explanatory considerations guide our inferences (Zumbo, 2007a, 2009). Explanation acts as a regulative ideal; validity is the explanation for the item or test score variation, and validation is the process of developing and testing the explanation. Zumbo (2009, p. 69) describes validation as an instantiation of an abductive method when he states that it is a higher-order integrative cognitive process involving every day (and highly technically evolved) notions like concept formation and the detection, identification, and generalization of regularities in data, whether numerical or textual. From this, understanding and explanation come after a balance of possible competing views and contrastive data.

As Stone and Zumbo (2016) argue, perhaps, as some hold (e.g., Borsboom et al., 2004), there are real, unobservable attributes that determine the performance, attributes that we are able to observe and directly measure, a performance such as responses in a mathematics achievement test or an assessment of intellectual functioning. Of course, such causal attributes may be embedded in a nomological net (Cronbach & Meehl, 1955); by assessment, neither Loevinger, Messick, Embretson, Zumbo, nor Schaffner preclude this possibility. I am unsure of Haig’s (in press) final position, but Borsboom et al. (2004) rule this out most certainly.

As an explanatory model of test score variation, Zumbo’s explanation-focused view of validity is embedded within an ecological model of item responding that is situated within a pragmatic view of abductive explanation wherein one develops validity evidence for tests through abductive reasoning (Stone & Zumbo, 2016; Zumbo, 2007a, 2009). In contrast to inductive or deductive reasoning, abductive reasoning neither construes the meaning of the scores purely from empirical evidence nor presumes the meaning and interpretation of the test to explain the score. Rather, abductive reasoning seeks the enabling conditions under which the score makes sense.

In my view of validity and validation, the explanans are elements of my ecological model (Zumbo, 2007b), which may be involved in setting the initial conditions of my abductive method. The item responses or test scores are the explanandum. In my explanation-focused view, my ecological model’s constituent concepts and variables (i.e., the explanans) explain the item responses or test scores (i.e., the explanandum). The role of the ecological model of item responding is described in detail in a subsequent section of this essay.

Contrasting with Kane's and others' argument-based approaches, perhaps the key distinction between an argumentation approach to validation and my explanatory approach is that the explanatory-focused approach is premised on developing validity arguments and switches the focus to how we decide which is the best argument or the best explanation.

Notably, I do not take as a first principle that the hypothetical construct (Cronbach & Meehl, 1955) or the latent variable (Borsboom et al., 2004) as a mapping of the empirical phenomenon explains the test score variation. The latent variable, or construct for that matter, may have explanatory value in some assessment settings, but this is not an essential part of my view.

In contrast to my view, reflecting the dominant empirical realist philosophy of the time, Cronbach and Meehl (1955) write:

Construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not "operationally defined." The problem faced by the investigator is, "What constructs account for variance in test performance?" (p. 282)

Determining what psychological constructs account for test performance is desirable for almost any test. (p. 282)

Loevinger (1957) adds an important level of nuance to the discussion when she persuasively argues that two basic contexts for defining validity should be recognized: administrative and scientific that play an important role in considering what needs to be explained (explanandum) and that which contains the explanation (explanans) in her validity theory. According to Loevinger, there are essentially two kinds of administrative validity: content and predictive-concurrent, whereas there is only one kind of validity that exhibits the property of transposability or invariance under changes in an administrative setting, which is the touchstone of scientific usefulness: construct validity (Loevinger, 1957, p. 641).

In other words, gathering test validity evidence during test design and development in a laboratory or controlled setting for use in the intended context(s) and population(s) where the focus is content and predictive-concurrent validity evidence. Setting aside Hempel's (1965) contentious view that adequate predictive arguments are potentially explanatory, neither of these forms of validity evidence has an explanatory aim, and Loevinger suggests that one is unnecessary. On the other hand, Loevinger's scientific context of test validity and assessment evidence drawn from the diverse and varying contexts of assessment use is where "[t]here is only one kind of validity which exhibits the property of transposability or invariance under changes in administrative setting which is the touchstone of scientific usefulness: that is construct validity" (Loevinger, 1957, p. 641). Loevinger states that, similarly to Cronbach and Meehl, the test performance is the explanandum that needs to be explained by the constructs (explanans). However, in her validity theory, Loevinger (1957) made the crucial point that every test, if for no other reason than the fact that it is a test, underrepresents its construct to some extent and contains sources of irrelevant variance; therefore, Loevinger may be the first validity theorist to open the door to the investigation of other constructs than the one being purportedly measured by the test in explanatory modeling of test performance. This notion is reflected in what I describe as the many ways of being human.

Regarding explanatory purposes, Zumbo et al. (2023) describe the importance of Embretson's groundbreaking research program, in which, in our terminology, the item responses are the explanandum (what needs to be explained), and the explanans contain elaborated cognitive models and componential decomposition include the explanation in her item response models of item response processes in support of test design and validation.

As we see in the quotations below, Borsboom et al.'s (2004) insistence on the explanatory power of the latent variable is foreshadowed by Cronbach and Meehl.

There is an understandable tendency to seek a "construct validity coefficient." A numerical

statement of the degree of construct validity would be a statement of the proportion of the test score variance that is attributable to the construct variable. This numerical estimate can sometimes be arrived at by a factor analysis, but since present methods of factor analysis are based on linear relations, more general methods will ultimately be needed to deal with many quantitative problems of construct validation. (p.289)

Rarely will it be possible to estimate definite "construct saturations," because no factor corresponding closely to the construct will be available. One can only hope to set upper and lower bounds to the "loading." (p. 289)

Borsboom et al. treat this explanation, in their view, as a causal explanation. A plausible empirical translation of their theoretical suppositions could be described as a literal reading of the arrows in a conventional path diagram of the factor analysis model as causal; that is, the latent variable is the causal explanation of the observed item response scores.

As Stone and Zumbo (2016, pp. 570-571) state, it should also be noted that the notion that constructs are unobservable entities determining observable actions is not generally accepted among validity theorists (see Slaney & Racine, 2013, for discussion), nor was this characterization of constructs posited as more than a possibility by Cronbach and Meehl (1955). Cronbach and Meehl also recognized that constructs emerge in collaborative inquiry practices. Construct validity, they noted, depended on the degree of agreement among researchers, which depended on the specificity of the theory or nomological net articulated by a construct's proponents.

Stone and Zumbo continue their analysis, stating that validating an assessment by utilizing constructs or causal attributes as the explanandum for a test score is fundamentally a pragmatic endeavor, depending on data, warrants, backing, and, finally, assertions that are testable and consistently useful. In this instance, pragmatism refers to the philosophic view. On the one hand, Borsboom et al.'s (2004) argument for causal attributes depends on their specification through the practices of measurement. On the other hand, as Cronbach and Meehl (1955), Kane (2013), and Zumbo (2007a) observe, construct validity depends on the development of an extensive, well-supported argument. Even then, construct validity may not be the best possible explanation for a test score. In language assessment, for example, time spent studying a language, how a person uses a language daily, whether a person uses that language at work, and other such factors may offer alternative competing explanations, as reflected in Zumbo et al.'s (2015) ecological model of item and test responding. In short, as both Kane and Zumbo have recognized, construct validity can play a role in developing the validity argument for an assessment. Still, it may not be the only role.

#### **4. SETTING THE STAGE FOR MY EXPLANATION-FOCUSED VALIDITY**

This essay section sets the stage for a detailed consideration of my explanation-focused validity by describing the confluence of ideas that influenced the development of my definition of explanation-focused validity and the aligned validation methods.

##### **4.1. What Motivated the Development of My Explanation-Focused View?**

At this point in the essay, it bears repeating that the description of my current theory is an explanation-focused validity that trends away from routine procedures toward an ecologically informed in vivo view of validation practices that embrace the many ways of being human.

The motivating factors for a novel validity framework are described in this sub-section of the essay to help assessment researchers consider the potential added value of a novel approach; we learn about the explanation-focused view by describing some of the reasons for its development. I developed the explanation-focused view of assessment research and validity theory because I was dissatisfied with test validity in the mid-1990s for the following reasons.

#### ***4.1.1. Avoid conflating test validity and validation: Developing innovations in test validation that derive from or require a particular definition of validity***

The first reason, as we saw in the historical analysis in the second section of this essay, is that several approaches did not clearly describe or define the concept of validity they were advocating. This lack of a definition or description may have been because some authors conflated test validity and validation; for example, validity is a correlation coefficient. In other cases, the definition of validity did not entail any particular validation method, such as a test is valid if it measures what it is supposed to. A consequence is that validation methods appeared ungrounded, lacking clear purpose, and incoherent. In contrast, I wanted a framework to develop innovations in test validation that derive from or require a particular definition of validity.

To make this concern less abstract, consider Messick's test validity theory. For the most part, even thoroughly expansive and systematic views of validity, like that of Messick, remained silent about a precise definition. However, to be fair to Messick, he either implied or acknowledged the importance of the earlier work on construct validity by Cronbach and Meehl (1955). For example, Messick (1995) describes the conventional view (content, criterion, construct) as fragmented and incomplete, especially because it fails to consider evidence of the value implications of score meaning as a basis for action and the social consequences of score use. He did highlight, however, that validity is not a property of the test or assessment but rather of the meaning of the test scores.

Regarding the absence of a description of the concept or a definition of validity, Shear and Zumbo (2014) show how this has had a trickle-down effect on the genre of reporting validity studies in educational and psychological research in academic journals. They state that without a guiding validity theory, assessing the success of validity research programs and comparing findings across different studies due to varying objectives is challenging. It bears repeating that in my view, in terms of the validation process (as opposed to validity itself), the statistical methods, as well as the psychological and more qualitative methods of psychometrics, work to establish and support the inference to the best explanation. This best explanation is "validity" itself, so validity is the explanation. In contrast, the validation process involves myriad methods of psychometrics to establish and support that explanation.

#### ***4.1.2. Bringing context back: Interpretation of test scores and the role and functions of assessment in society***

The second reason for my dissatisfaction with the state of affairs in validation practices in the mid-1990s reflected a mostly uncritical acceptance of context-free interpretations of scores from tests, measures, and surveys. In a parallel line of research with Donald Zimmerman, we continued the development of a mathematical framework he introduced in 1975 in *Psychometrika* for mental test data (Zimmerman, 1975). I have come to call this abstract mathematical framework "measure-theoretic test theory" or "measure-theoretic mental test theory," which provides a more rigorous description of classical test theory (CTT) founded on the notion that the data we observe arises with a particular type and amount of uncertainty reflected in the generic statement  $X = T + E$ .

Ultimately, measure-theoretic test theory liberates us from the received view of the true score as immutable and unchanging. It allows us to re-interpret the true score as contextualized, situated, and ecologically shaped. This re-interpretation of the true score closely aligns with the critical components of my explanation-focused view of test theory, validation practices, and assessment research. I describe this development of the re-interpretation of the true score in  $X = T + E$  in a subsequent section of this essay.

While co-chairing with Suzanne Lane the technical working group in support of the United



States of America's Congressional review of the National Assessment of Educational Progress, NAEP (Lane et al., 2009), my view of the role and functions of assessment in society and the school system was solidified. The impact of social and cultural issues at the system macrostructure and the classroom microstructure can be seen in my centering on the role of values, consequences, and the many ways of being human in test validation (Zumbo, 2018a) and in developing a multilevel test validity theory (Zumbo et al., 2017; Zumbo & Forer, 2011) and reflects yet another implication of bringing the context back into psychometric test theory (Zumbo, 2009). I unpack this in a subsequent section of this essay related to values, context, consequences, and the many ways of being human.

#### ***4.1.3. Dissatisfaction with context-free models of explanation and hypothetico-deductive methods***

Third, developments in the philosophy of science and test validity related to educational research on learning, achievement, and human development, along with psychological inquiry into traits, dispositions, and attitudes of the imperative of a contextualized view of the phenomena that did not align with dominant views of test validation by Cronbach and Meehl (1955).

Cronbach and Meehl's logical empiricist view of the nomological network's commitment to the covering law account of explanation Zumbo (2009) and the hypothetico-deductive theory of confirmation (Haig, in press). The covering law account of explanation and the hypothetico-deductive theory of confirmation was considered *de rigueur* in the philosophy of science around the time, and shortly after, Cronbach and Meehl (1955) introduced construct validity. However, over the 70 years, many concerns have been raised, and they are no longer the dominant views.

As Zumbo (2009) noted, the most critical problem with Cronbach and Meehl's nomological network approach is that it attempts to characterize explanation as context-free, like its covering law forefather. Zumbo (2009) and Stone and Zumbo (2016) criticize the covering law model of explanation in test validity because, from their vantage points, an explanation is a "pragmatic" or "contextual" concept-- an idea that the covering law models and their variants seem to reject. On a related note, in the seven decades of philosophical inquiry, since the covering law model was introduced, the large body of research literature in the philosophy of science focused on explanation can be characterized by the search for an explication of the locution "scientific explanation" and for the construction of powerful explanatory models. However, this development, for the most part, kept physics as the reference science. That is good and fine for physics, but educational and psychological assessment and testing are substantially and nontrivially different from physics in terms of their theory structure and development and functional status. As such, an explanatory model for educational and psychological testing and assessment should be informed by the scientific method in the psychological, educational, and behavioral science offered by methodologists such as Haig (2005b, 2014, 2018, 2019).

Early in developing my explanatory view (Zumbo, 2007a, 2009), I made the case that validity is a matter of inference and the weighing of the evidence in explanation-focused theory. I also noted that explanatory considerations guide our inferences; construct validity centrally involves making inferences of an explanatory nature and emphasizes the importance of explanation as a pragmatic endeavor. Moreover, our construct validation efforts should be guided by explanatory considerations in which the goodness of our explanatory theories is assessed by a process of inference to the best explanation.

Stone and Zumbo (2016) contribute to the explanation-focused view by, in good part, addressing how contemporary assessment practitioners, researchers, and educators can utilize the strengths and minimize the shortcomings of a science of measurement informed by pragmatic concerns. They describe, among other things, how a certain American pragmatism—

as articulated in works of such philosophers as Williams James, John Dewey, and Charles Sanders Peirce—provides a framework in which to approach critical foundational issues in test validity to begin to break down the wall dividing scientific practice and theorizing about the concepts of validity. Pragmatic explanatory methodology in assessment and testing aims to embrace justice and fairness (which I describe within the concept of the many ways of being human) by respecting practical, pluralistic, and provisional dimensions of pragmatic explanation.

#### ***4.1.4. Taking the value-laden stance further by bringing what I describe as the many ways of being human into the foreground***

The fourth reason for my developing the explanation-focused approach was to create a validity theory that fostered an attitude among assessment researchers to embrace the many ways of being human.

In a subsequent sub-section of this essay, I make the case that Messick's (1980, 1989, 2000) theoretical developments in a validity theory that viewed values and consequences as an integral part of construct validity and the validation process as they contribute to the soundness of score meaning, were nearly concomitant with developments in the philosophies of science that began to consider a value-laden stance that guides epistemic integrity. I wholly concur with Messick's developments along this line of reasoning and aim to take the value-laden stance further by bringing into the foreground what I describe as the many ways of being human that aim to inform validation practices from their initial planning. I believe this aligns with Messick's view of the role of values and consequences and opens further the discourse of validity evidence that will encourage us to shine a light on hidden invalidities.

#### ***4.1.5. Emphasizing the importance of response processes***

The third reason for developing the explanation-focused view is that it allows me to influence assessment research more generally and validation research in particular, emphasizing the importance of response processes and embracing the many ways of being human in the design and interpretation of the findings.

As Shear and Zumbo (2014) describe it, by the year 2000, researchers reporting validity studies in many educational and psychological measurement journals commonly included more diverse evidence to support test score interpretations than they did in the mid-1960s and 1970s, with notable increases in factor analytic and content-based evidence. However, validation research has continued to leave out validity evidence based on the response processes of examinees and the consequences of test use. In addition, although researchers seem to consider more (and more complex) sources of evidence, clear theoretical bases for such practices, such as the concepts of validity described above, were not explicitly stated.

## **4.2. Context, Ecology, Diversity, and the Many Ways of Being Human**

The arguments motivating the importance of context, ecology, and the many ways of being human begin with the recognition that embodied or distributed cognition is present when a respondent or test-taker encounters a task or item on a test, assessment, or survey. I have been persuaded of the importance of bringing Varela et al.'s (1991) description of the embodied mind and, more broadly, contemporary notions of distributed cognition, such as those of Clark's (1998), into assessment and testing research. In broad strokes, these views of cognition reflect a circulation between cognitive science and human experience, fostering the possibilities of human experience in a scientific culture of assessment and testing research.

However, suppose something like this embodied or distributed view of cognition is correct. How does this generally affect our conceptualization and practice of test validity, validation research, and assessment research? The response to this question has two parts. The first part

signals the importance of the testing situation or context and diversity of the test takers, as Zumbo et al. (2023) state:

To take “embodied” seriously means to consider their neurological and chemical basis, as well as the social and ecological significance of context and the “extended mind” (Clark, 2011), whether it involves virtual phenomena in onscreen interactions or the wider significance of the testing situation (e.g., setting, time, stakes).

To further explore these themes, we will consider the significance of disability and neurodiversity in tested populations. There is a broad diversity within human neurobiology (Pellicano & den Houting, 2022); the human brain develops and functions in countless ways, resulting in a test-taking population with diverse strategies and responses. There is a need to recognize that, rather than anomalies, test-takers with disabilities and learning differences represent a sizeable minority. (p. 257)

Zumbo et al. (2023, p. 255) continue this line of reasoning and argue that response processes to test items or tasks involve the cognitive strategies and approaches of test takers and emotion, affect, interaction, physiology, and embodied behavior in the test ecology. In my view, as described in Zumbo (2015), what I refer to as *in vivo* (as opposed to *in vitro*), the context is not a nuisance that “distorts the picture” but instead informs and shapes the attributes—i.e., one cannot extract the context.

This *in vivo* view is reflected in Zumbo et al.’s (2015) description of their ecological model of item responding, wherein contextual factors could affect item responses by mediating the cognitive processes that are usually assumed to generate item responses. In so doing, as they state, they accept as the starting point of the argument the widely received view in the broader social sciences that test takers bring their social and cultural present and history to test taking and that human beings have evolved to acquire culture from birth, and that the culture to which an individual is exposed, and the ecology of their lives, affects their basic psychology and cognition, including, in our case, item responding. In so doing, one can move to a contextualized form of explanation that works against a binary structure of variables that explain test performance (Zumbo et al., 2015, p. 140).

From a psychometric perspective, this *in vivo* view is based, in large part, on our developments in measure-theoretic test theory (Kroc & Zumbo, 2020; Zimmerman & Zumbo, 2001; Zumbo & Kroc, 2019) interpretation of a true score. Furthermore, from a theory of validity as social practice, Addey et al. (2020, p. 588) address the question: How should different validity arguments and evidence be reconciled in situations where there are diverse stakeholders and multiple contexts of use?

The concepts described above come together to reflect a central idea in the current essay: “the many ways of being human,” which reflects the diversity and complexity of the human experience. It acknowledges numerous ways to live, think, express, and experience the world as a human being, encompassing many aspects, including but not limited to cultural practices, personal beliefs, emotional experiences, and physical realities. Therefore, the centrality of the many ways of being human, as embraced in my explanation-focused view, can be seen as a celebration of this diversity and a call for assessment researchers to explore and understand the breadth and depth of the human condition.

However, it is critical to note that the interpretation of this phrase can vary based on context and individual perspective. Some might see it as a philosophical question about the nature of humanity as it relates to testing and assessment. In contrast, others might view it as a call for empathy and understanding in recognizing how people live their lives. Therefore, embracing the many ways of being human must be more than a performative act of our collective desire toward fairness and inclusion in testing and assessment practices and research. These many ways of being human also need to be more than just an ambition beyond our collective grasp

and more than a regulative ideal. The many ways of being human need to shape and inform our research at the core of our methods, including the importance of consequences and values in testing and assessment, as will be demonstrated in the section of this essay focused on innovations in methodology.

#### **4.3. Recognizing and Quantifying Uncertainties in Test Validation and Assessment Research Practice**

Uncertainty is ubiquitous in science, but scientific knowledge is often represented in the public and policy-making contexts as certain and immutable (see, for example, Giere, 2010; Gigerenzer et al., 1989). Ignoring uncertainty can foster distrust in assessment research when they are derived in a way people perceive as pernicious and arbitrary, making it inadmissible. For this reason, the quantification of uncertainty is reflected in the theoretical developments of our validity framework and the methodological innovations described later in this essay.

Consider, for instance, the uncertainty due to the variability in performance on a test that may be due to factors such as familiarization of the test delivery modality, for example, computer-based administration, pacing, or calibration of instruments. This uncertainty is widely discussed in educational and psychological measurement because tests or assessments cannot measure the phenomenon they purport to measure perfectly. This uncertainty travels under the umbrella term “measurement error” in educational and psychological measurement. Far less widely known is that six additive measurement error models are deceptively similar in their general algebraic form,  $X = T + E$ , but have different error structures that connect and distinguish them (Kroc & Zumbo, 2020). look commonly used in disciplines from psychometrics and test theory to economics to epidemiology.

Loevinger (1957) made the crucial point that every test if for no other reason than the fact that it is a test and not a criteria performance, underrepresents its construct to some extent and contains sources of irrelevant variance. As such, it is important to distinguish two additional forms of uncertainty.

- The first additional form of uncertainty is its central role in statistical models that result in probabilistic statements about the world.
- The second additional form of uncertainty is characterized by its central role within explanatory theories, for which models take the form of probabilistic claims about the world (Gigerenzer et al., 1989).

Negotiating these and other forms of uncertainty through constructively arguing and presenting a transparent and logical case building toward consensus agreement while uncertainty is present is a crucial part of the scientific process (Giere, 2010; Gigerenzer et al., 1989).

As described by many methodologists and philosophers of science going back to the early part of the last century, science is a process that builds better models which increasingly allow us to make increasingly more accurate theoretical and empirical predictions (for example, Carnap, 1935; Giere, 2010; Lakatos, 1976; Reichenbach, 1977). This process is crucial to recall in all assessment research, particularly test validation research. To make this less abstract, let us consider social and personal consequences and side effects (Hubley & Zumbo, 2011) for a case of tests that lead to a pass/fail decision, entry into college, or licensure. For example, recognition of the region of uncertainty around the cut-score and purported impact and negative consequences and proactive policies emerging from the definitions of negative impact to deal with findings that fall in that region diminish the likelihood of false-positive (a claim regarding the impact of negative consequences effects when they should not) and false negative (a claim of no impact of these adverse effects when they should) results. There are potentially severe consequences to both false outcomes. Understanding systematic and random variability, the size of the region of uncertainty, and developing appropriate policies to deal with such findings

results are fundamental to best practices informing defensible test validation research.

Although there is a history of considering and quantifying this uncertainty as measurement error going back to the early 1900s, we will see in the section below that recent developments in the mathematical structure of that test theory were significant in defining my explanatory focus on the variability of item responses and sub-test or test scores guided by shaped by the ecological model of as defining features of test validity and shaping validation practices. In a subsequent section, we will see that this contrasts with other views of validity, where the source of the explanatory focus is on the construct theory or latent variable.

#### **4.4. Initially, Classical Test Theory Seems Simple, but Its Description and Interpretation Have Changed Over Time and Is Now Aligned with the Explanation-Focused View**

##### ***4.4.1. Classical Test Theory (CTT) has been the source of tremendous innovation and generated much confusion***

Spearman's (1904) characterization of an observed score as a sum of a true score and an error was responsible for tremendous development and innovation in what has come to be widely referred to as CTT applies to any measurement process, including, for example, educational tests, psychological instruments, and observation ratings based on rubrics or checklists, to name a few. In their most common use in assessment and testing, a defining feature of these various examples of a measurement process is classical test theory's focus on the individual test-taker, study participant, or survey respondent. CTT applied to mental tests has a long history of application to test construction, psychometric analysis, and utilization of technology for test delivery. As Raykov and Marcoulides (2016, p. 325) state, "[f]or much of the past century, classical test theory (CTT) was the dominant framework for developing multicomponent measuring instruments in the educational, behavioral, and social sciences." Nonetheless, not long after Spearman's initial description in 1904, it generated much confusion and controversy among psychometricians, educational and psychological assessment specialists, and researchers.

Of particular importance for test validity and my explanation-focused view of validity and assessment research more broadly is the nature of the true score. To my knowledge, Raykov and Marcoulides (2011, Chapter 5) provide the most thoroughgoing description of common misconceptions of classical test theory and their correct interpretation in the psychometric literature. It is accessible to applied researchers and assessment specialists.

##### ***4.4.2. What do we mean by "Classical Test Theory (CTT)"?***

To avoid confusion, I must explain that I use the phrase "classical test theory (CTT)" throughout this essay to describe a theory involving three canonical concepts of an observed test score,  $X$ , which stands in for the unobserved true score,  $T$ , and the measurement uncertainty reflecting a discrepancy between  $X$  and  $T$  denoted  $E$ .

- Quite correctly, the burgeoning discipline of individual differences psychology is often described as the progenitor of the description of psychological and educational measurement uncertainty as an additive error by the generic statement  $X = T + E$ . As such, the model that travels widely under the name "classical test theory" can be considered a legacy of Spearman (1904).
- It is worth noting that other disciplines have had their concerns about measurement uncertainty. As such, Kroc and Zumbo (2020) describe five additive error models commonly used in disciplines from psychometrics and educational assessment and testing to economics to epidemiology and one new model formerly proposed in Kroc & Zumbo (2018). These models share the general algebraic form,  $X = T + E$ , but have different error structures that connect and distinguish them.
- The psychological measurement error model was among the first and was unique in that, for



the most part, psychological researchers at the turn of the 1900s were interested in the uncertainty evidenced at the between-person level, which was unsurprising given the interest in empirical studies of the sources and reasons for individual differences. This individual difference model sat well with and also became widely used by psychologists and educationalists interested in the role of measurement uncertainty in assessing individual students or clients in mental health settings.

- The focus herein is on the mathematical structure of the model and not on estimation or inference. As such, the description of CTT does not require any particular distributional structure to the error terms beyond the primary exchangeability conditions described in Kroc and Zumbo (2020). In particular, no parametric assumptions are required of the CTT model at the level of mathematical abstraction I use here.
- Finally, estimation or inference with CTT will require additional assumptions. I will provide two examples with slightly different foci. In the first example, if one were interested in using CTT when specifying specific latent variable statistical models such as factor analysis to investigate and quantify sources of between and within-person variability with likelihood theory estimators from repeated measures data. A second example reflects a different use of the CTT model herein, where the classical mathematical object of test reliability derived from the CTT model requires that both the true score and the error be square-integrable (Zimmerman & Zumbo, 2001). This additional assumption is not required of the original CTT model. However, it is crucial in the inferential framework for the classical test theory.

#### ***4.4.3. Informal, classical, and measure-theoretic periods, each of which resulted in a mental test theory model that is representative of that period***

I have used “measure-theoretic test theory” without defining it. I will define it in this section by contrasting it to test theory derived during the informal, classical, and measure-theoretic periods of development, each resulting in a test theory model representative of that period.

In short, however, measure-theoretic test theory uses the language and concepts of measure theory and probability spaces to describe the axioms of mental test theory. In contrast, if the reader is sufficiently well-versed in measure theory or measure-theoretic probability, Lord and Novick’s (1968) mathematical descriptions suggest measure-theoretic concepts (i.e., measure theory, if you wish, can be read between the lines). However, their theorems and principle results are not expressed using measure theory, likely attributable to their intended audience of psychological researchers and psychometricians (Kroc & Zumbo, 2020).

I will describe three periods of theoretical development of test theory models: the foundations of the latter two are grounded in statistics, probability or measure theory, and functional analysis. The adjectives “informal,” “classical,” and “measure-theoretic” will be used to describe a specific genre of inquiry or the language used in developing and describing the CTT model in these three developmental periods.

The three adjectives were also chosen because they reflect the similar historical development of informal, classical, and measure-theoretic probability theory concepts. However, advanced study and rigorous descriptions of probability consider it a branch of mathematics and typically necessitates measure theory. Notably, although there are no standard descriptors of probability used in all disciplines, there are widely used normative practices under which I am using the term “classical probability” in a boutique manner to allow the comparison with test theory. The critical point is that measure-theoretic probability has a distinct feature of using the language and concepts of measure theory, which the other two do not. The same distinction holds for test theory. As such, I acknowledge that there may be some confusion from using the term “classical” to refer to both a test theory model statement (i.e.,  $X = T + E$ ) and a period in reflecting the development of the CTT model; therefore, I will mark the latter by the phrase “classical period.”

Gulliksen (1950b), Guttman (1945), Lord and Novick (1968), Novick (1966), Rozeboom (1966), and others are representative of the classical period in test theory, which explicitly defined observed scores, true scores, and error scores as random variables, having designated properties. These formulations improved on the less systematic formulations of what I refer to as informal test theory that had prevailed earlier in the century. It is worth noting that some writers used to describe developments in the informal period. However, when it was used during the informal period, it was less rigorous formalism than seen during the classical development period. The CTT model, as described in Lord and Novick (1968) and formalized by Zimmerman (1975), proposes that each respondent has a fixed true score,  $T$ , capturing the attribute of interest. The classical period in test theory derives from the pioneering work of Spearman and Yule, which is summarized by Gulliksen (1950b). Zimmerman (1975) is the landmark paper that signaled the beginning of the measure-theoretic period in test theory.

In 1966, Melvin Novick published a landmark paper entitled “The axioms and principal results of classical test theory” that, in an important sense, started the process toward measure-theoretic test theory. Novick motivates his work by describing how the model of test theory dominant in the classical period “... suffers from some imprecision of statement so that, from time to time, controversies arise that appear to raise embarrassing questions concerning its foundations” (Novick, 1966, p. 1). A little over a half-century after Novick’s statement, Kroc and Zumbo (2020) document classical test theory mischaracterizations found in the recent work of psychometricians and applied measurement specialists. Calling for further analysis of test theory models and a description of the connections between six linearly additive measurement error models that are variations of  $X=T+E$ , they state: “The need for such clarity becomes apparent when one reviews the classical test theory (CTT) literature, which is littered with false characterizations of its measurement error model” (Kroc & Zumbo, 2020, p. 1).

Therefore, Novick’s (1966) axiomatization of the classical period signaled an essential change in the development of the models in the classical period. For most purposes, identifying test scores with random variables is all that is needed to develop the theory and make the mathematics of probability and statistics available. However, the distinctive character of test theory and its relationships with other mathematical models becomes more evident when incorporated into an abstract mathematical framework using measure theory.

Two features of CTT are described as a demonstration of this distinctive characteristic of CTT that has stimulated much debate in psychometric research. First, the CTT model described by Novick (1966) and described in greater detail in Lord and Novick (1968) is representative of the classical period, focused on measurement error, and as described in Zumbo and Kroc (2019), among others, invokes a type of hierarchical structure, and a hypothetical propensity distribution for each test-taker, the expected value of which is that test taker’s true score. In yet another case of expository metaphor running amok when describing nuanced mathematical ideas to an audience not all of whom have sufficient mathematical preparation, this propensity distribution is often described, as it was by Lord and Novick, as a random variable with a distribution over imagined replications of the test with the test taker’s memory wiped between replications.

Second, notably, Novick and Lord used random variables (and their attendant properties) to model probabilistic concepts in mental test theory rather than actually be the concepts themselves. This distinction is implicit in much of psychometric theory when we distinguish between an abstract version of a mathematical object and a concrete representation (or model) of that object. Therefore, these authors and others who followed by using the memory-wiping metaphor (a type of concrete representation) to describe the more nuanced mathematical object of a true score, such as the probability distribution of a conditional random variable (i.e., the propensity distribution) that represents the inherent variability, or error of measurement,

characterizing a person's test score. In this case, the abstract version of the mathematical concept is correct; however, outside of films wherein people's memories are supposed to be wiped (see the "Men in Black" series of movies), the concrete representation is nonsensical and potentially misleading readers to accept as given notions like the necessity for parallel tests, and strong conditions such as experimental or local independence. At the same time, overshadowing a unique feature of CTT compared to other error models (Kroc & Zumbo, 2020) that Zumbo and Kroc (2019) and others show: that the definition of the true score assures that each test-taker or survey respondent receives one and only one true score that remains fixed on any actual or hypothetical reapplications of the measurement process  $X$ .

#### 4.5. Some Remarks on Measure-Theoretic Test Theory

Measure-theoretic test theory aims to describe the properties of test theory related to the theory of properties of conditional expectations of random variables defined on probability spaces was initiated by Zimmerman (1975) and continued by Steyer (1988, 1989), Steyer and Schmitt (1990), and recent developments investigating various error models of which the prominent test theory (classical test theory) model in an instantiation by Kroc and Zumbo (2020). Zimmerman and Zumbo (2001) considered test theory from the perspective of measure theory on Hilbert spaces, showing that the higher the level of abstraction, the more comprehensive the unification of diverse interpretations of test theory.

##### 4.5.1. Measure-theoretic mental test theory: CTT

As Zumbo and Kroc (2019, p. 1187) state, the classical test theory (CTT) model, as described, for example, in Lord and Novick (1968) and formalized by Zimmerman (1975) and described in more detail below, proposes that  $X = T + E$ , where  $\mathbb{E}((X|\sigma(f)))$ , where  $f$  is an assignment-to-individuals function and  $\sigma(f)$  denotes the set of measurable events generated by this function. More details are provided in Kroc and Zumbo (2020), Zimmerman (1975), and Zimmerman and Zumbo (2001). Under the CTT model, the definition of the true score assures that each test-taker or survey respondent receives one and only one true score that remains fixed on any actual or hypothetical reapplications of the measurement process  $X$ .

Three equivalent formulations of measure-theoretic classical test theory follow; Kroc and Zumbo (2020) prove the equivalence of these three formulations of the CTT model in detail. Formally, this model is defined via a measurable space  $(\Omega, \mathcal{F})$  on which  $X$ ,  $T$ , and  $E$  are defined as real-valued random variables and an assignment-to-individual function  $f: \Omega \rightarrow \Phi$ . The image space  $\Phi$  is thought of as the space of test-takers or survey respondents; thus, for any individual  $\phi \in \Phi$ , we construe  $X(f^{-1}(\phi))$  to capture all possible outcomes of the measurement process  $X$  for the particular individual  $\phi$ . Let  $\sigma(A)$  denote the usual  $\sigma$ -algebra generated by the generic function (or random variable)  $A: \Omega \rightarrow \Lambda$ ; i.e.

$$\sigma(A) := \{A^{-1}(S) : S \in \Lambda\}.$$

The classical test theory model described above can then be compactly expressed as follows (Zimmerman, 1975):

$$X = T + E, \text{ where } T := \mathbb{E}((X|\sigma(f))), f: \Omega \rightarrow \Phi. \quad (1)$$

The model was reformulated by Zimmerman and Zumbo (2001) as follows:

$$X - E \text{ is } \sigma(f)\text{-measurable, } T := \mathbb{E}((X|\sigma(f))), \mathbb{E}((E|\sigma(f))) = 0, \Omega \rightarrow \Phi. \quad (2)$$

Notably, Zimmerman and Zumbo's reformulation in model (2) does not a priori specify a functional relationship between the three canonical quantities  $X$ ,  $T$ , and  $E$ .

Kroc and Zumbo discuss the CTT model's properties regarding sample units' exchangeability. For the CTT model, the error terms must balance on the individual; this is the requirement that

the expected value of the error is zero over all possible measurements of each particular individual- i.e., individual-level exchangeability of errors condition. This condition is the key, novel structure of the CTT model; without it, we would not have the defining property that the expectation of the observed score should equal the true score for every individual (see Kroc & Zumbo, 2020, for more discussion).

More than one plausible sample space may be available, depending on the assessment design and setting. Although more complex cases are described later in this essay, the simplest case involves items and test takers, which may be constructed as the Cartesian product of the two (or more) sample spaces. As Zimmerman and Zumbo (2001) note, formally, test data are the realization of a stochastic event defined on a product space  $\Omega = \Omega_I \times \Omega_J$  where the orthogonal components,  $\Omega_I$  and  $\Omega_J$ , are the probability spaces for items and examinees respectively. The joint product space can be expanded to include other spaces induced by raters or occasions of measurement, a concept formalized in generalizability theory. Hence, modeling test data minimally requires sampling assumptions of a hierarchical experiment (i.e., measurement process) about items and examinees and the specification of a stochastic process that is supposed to have generated the data.

#### 4.5.2. Function spaces, metric spaces, and Hilbert spaces

Zimmerman and Zumbo (2001) introduced an operator theory formulation of CTT by describing the measurement process as a collection of linear operators acting on a Hilbert space of true score vectors. This way, true and error scores can be naturally associated with projection operators on this Hilbert space. Once this identification is made, metric concepts of distance, length, angle, and orthogonality have immediate implications for test theory. They went on to show, exploiting their operator formalism, that one can consider reliability as a mathematical object that can be defined as another type of projection.

The collection of all observed scores associated with a measurement process represented by the function space

$$L^2(\Omega, A, P);$$

the collection of all true scores is the Hilbert subspace

$$L^2(\Omega, B, P), B \text{ is a } \sigma\text{-algebra contained in } A.$$

Moreover, the collection of error scores is the orthogonal complement of the subspace of true scores.

Notably, it is not necessary to consider the collection of all random variables defined on a probability space to interpret concepts in probability, statistics, and test theory. It is sufficient to restrict attention to the collection of all random variables having finite variance, or, as sometimes called, square-integrable random variables. Because random variables with finite variances also possess finite covariances and expectations, this collection is sufficiently large to provide for an interpretation of test theory.

Zimmerman and Zumbo define the true score as a linear operator acting on random variables and the error score as a linear operator. The collection of all true score random variables, or B-measurable random variables, are defined on the same probability space.

This probability space is a Hilbert subspace of the space of observed score random variables. The distinctive features of test theory as a mathematical model are closely related to the fact that the true score operator is a projection operator in Hilbert space. Therefore, the conceptual definition of CTT reliability is equal to one if and only if the observed score random variable equals its corresponding true score random variable (Zimmerman & Zumbo, p. 290).

From this formalization, a reliable test score is one that is “close” to the subspace of true scores

so that the length of its projection is almost the same as its own length. Such ideas are familiar in least-squares regression. Suppose the length of the projection is decidedly less than that of the original vector. In that case, the two are “almost” perpendicular so that reliability is close to zero. Along the same lines, the reliability of a test can be regarded as the “Rayleigh quotient” of an observed score centered at its expectation with respect to the true score operator.

Extending this reasoning further, Zumbo (2007a, p. 74), building on the connection described above to regression and a geometric partitioning of the regression model R-squared (i.e., the Pratt index), argues that one can consider the generic measurement model statement  $X=T+E$ , on par with the generic regression model statement described in Zumbo (2007a, pp. 66-69). Apply the geometry in Zimmerman and Zumbo (2001). One can show that classical test reliability is, in essence, a Pratt index – a partitioning of the explained variation in the test score attributable to the model, just like an R-squared value in regression.

It is well known that a conceptual definition of the classical test theory reliability is the squared correlation between observed scores and true scores. Thus, a natural definition of the mathematical object test validity and a valid test score can be defined similar to a reliable test. This definition, however, is of limited value in the Novick or Lord and Novick description of CTT because the true score ignores the context or situation of the measurement process. On the other hand, the re-interpretation of true scores as an affordance of measure-theoretic test theory reminds us that discussing what it means for a test to be valid requires consideration of the context in which the test taker and measurement process are situated, in the manner similar to explanation-focused validity.

This interesting definition of validity does not involve the criterion (predictive or concurrent) validity description that sheds some light on the concept of validity and is a geometric interpretation akin to Cronbach and Meehl (1955) and Borsboom et al.’s definition, see sections two and three of this essay without the layer of construction and assumptions required of a latent variable model in their definition. Furthermore, this definition reminds us that because  $T$  is unobserved, there is little one can do about estimation and inference with this geometric description of validity, which is why I refer to it as a conceptual definition. Test theorists of a century ago were most certainly aware of this, which provides insight into the clever step of designing an experiment with a criterion variable to side-step the problem of the unobserved variables. Likewise, this conceptual definition highlights the importance of explanatory approaches to the item and test performance, where the item or test performance needs to be explained (i.e., explanandum). The ecological model of item and test performance provides a framework to consider what contains the explanation (i.e., the explanans).

#### **4.6. The Re-interpretation of the True Score of CTT is an Affordance of Measure-Theoretic Test Theory That is Important to My Explanation-Focused Validity and Assessment Research**

In this sub-section of the essay, I argue that (a) a re-interpretation of true scores, and hence observed scores, of measure-theoretic test theory that, unlike conventional interpretations of classical test theory (CTT) such as that of Lord and Novick (1968), allows for an ecologically shaped, in vivo, true and observed test score, and (b) this alternate re-interpretation provides the psychometric building blocks of a coherent explanation-focused approach to test validation and assessment research.

In short, measure-theoretic test theory allows for an alternate interpretation of CTT’s  $X = T + E$ . This new re-interpretation aligns with the description in a preceding sub-section of this essay that focused on the importance of context, ecology, and the many ways of being human, with the recognition that embodied or distributed cognition is present when a respondent or test-taker encounters a task or item on a test, assessment, or survey.



The alternate interpretation of true and observed scores reflects my view of the importance of context or situation in interpreting test or survey scores (Higgins et al., 1999; Zumbo, 2007a, 2007b, 2009; 2017), my developments of an ecological model of item responding and test scores (Zumbo, 2007b, 2009; Zumbo et al., 2015), the importance of distinguishing what I refer to as *in vivo* versus *in vitro* views of assessment (Zumbo, 2015), and trending away from routine procedures, toward with an ecologically informed *in vivo* view of validation practices (Zumbo, 2017).

It is worth noting that based on results in Zimmerman (1975) and Zimmerman and Zumbo (2001) using the language and methods of measure theory, both the conventional and re-interpret of the true score are allowable; however, the Lord and Novick (1968), and Novick (1966) model of CTT, only allows for the conventional interpretation of the true score.

#### ***4.6.1. Contrasting the conventional interpretation and the re-interpretation of the true score of test theory***

Let us focus on getting a deeper appreciation for the re-interpretation of the true score of CTT by contrasting the conventional interpretation to the re-interpretation in two assessment settings: one-point-in-time assessment and repeated measures assessment designs.

The various interpretations of classical test theory based on the Novick (1966) and Lord and Novick (1968) axiomatization and Zimmerman's (1975) axiomatization of  $X = T + E$  typically involve explaining the mathematical formalism and, perhaps, creating a mental or physical image of the theory. While the mathematical structure described by Zimmerman and extended by Zimmerman and Zumbo (2001) has a strong foundation and more adequate axiomatization that permits Novick and Lord's interpretation, there is still much to be resolved about its various interpretations. I wish to highlight that when used in the context of this section of the essay, "interpretation" is plural because, in many cases in advanced mathematics, abstract mathematical objects may have various cognitive or physical interpretations even if the mathematics. There are many examples of this in physics.

**4.6.1.1. Conventional Interpretation of The True Score of CTT.** The conventional interpretation of the true score is founded on the view that the true score is a property of the test taker. It is important to note that the interpretation of a true score as a property of a test-taker arises in the classical test theory formulations such as those of Guttman (1945), Lord and Novick (1968), and Novick (1966), where a true score was defined as the expectation of an individual's observed scores over independent, repeated measurements or replications of a test. Lord and Novick introduced the "propensity distribution" and an accompanying notation as a mathematical object characterizing a test-taker's hypothetical distribution of observed test scores arising from the memoryless replications of a test. By this interpretation, a person's true score is commonly defined as the expectation over an infinite number of independent test administrations. Thus, largely due to the "wiping the test taker's memory clean between replications," the variation in observed scores is due to measurement error for repeated measures.

It is important to note that the definition of true scores in the various models described in Zimmerman and Zumbo (2001), such as classical models described in Novick or Lord and Novick, the measure-theoretic models, including those that center on the conditional expectation, as well as the operator theory and Hilbert space models, are, from a mathematical perspective, all equally valid or true. However, some may be more useful or attractive than others. Therefore, choosing between the classical Lord and Novick model and the measure-theoretic models is a matter of interpretation.

That is, from a mathematical perspective, defining the score,  $T = \mathbb{E}(X|\sigma(f))$ , where  $f$  is an assignment-to-individuals function is fine. However, without measure theory, one must invoke

some version of a "wiping the test taker's memory clean between replications," which explains why Lord and Novick and others resorted to this in their descriptions. It also explains why many descriptions of CTT insist that it characterizes a repeated measures assessment experiment; after all, it is in the definition of the true score. This metaphor also explains why some writers describe CTT as imposing immutable outcome variables, why simple difference scores are treated as inherently poor measures of change (Zumbo, 1999), and why I describe this practice as a metaphor run amok.

**4.6.1.2. The Re-Interpretation of the True Score of CTT for a One-Point-In-time Assessment (Cross-Sectional Assessment Design).** In contrast to the conventional interpretation, the new re-interpretation of the true score one is seen as conditioning on all possible outcomes of the measurement process  $X$  for a particular test-taker or survey respondent. Suppose we imagine obtaining infinite observations from a test-taker in various ecological testing settings, denoted  $\mathcal{S}$ , of the sort described, for example, in the ecological model of item responding and test performance (Zumbo et al., 2015). In that case, the true score for test-taker  $j$  is the mathematical expectation of all observations over the varying ecological testing setting represented in  $\mathcal{S}$ . Therefore, the variation in observed test-taker scores includes measurement error and variation attributable to the different test ecological testing settings reflected in  $\mathcal{S}$ . Stated differently, the re-interpretation of the true score in the scenario of the various ecological testing settings, a test-taker's observed test score can change depending on the varying ecological testing settings represented in  $\mathcal{S}$ .

Kroc and Zumbo (2020) describe the exchangeability condition of the CTT model. Beyond the mathematical statement of CTT using measure theory, we described the model in the context of the designed assessment experiments reflected in the concordance setting where test takers are assigned to selects of  $\mathcal{S}$ , defined above. Alternatively, one may administer a measure similar to assessment practices in which a survey or instrument is administered in a less tightly controlled setting and test takers are not allocated to all or a subset of ecological settings in  $\mathcal{S}$ , defined above. As Kroc and Zumbo note, both the tightly controlled and less tightly controlled versions of the assessment design align well with generalizability theory governing principles that aim to understand measurement processes through an experimental design framework. Further semantic interpretation of a feature of CTT is described in Zimmerman and Zumbo (2001) in the language of measure theory and functional analysis, which is notable at this juncture is that it allows for different observed score distributions for test-takers with the same true score.

Two examples may help make this new re-interpretation of the true score less abstract. An operational example of this interpretation can be seen in Chapter 2, Section 2 of Zumbo (2021), wherein I describe the principles and logic of my methodology to investigate the concordance of various test delivery and administration settings in online computer-based testing. That is, I use Zimmerman and Zumbo's (2001) measure-theoretic (Hilbert space) approach extended to outline the methodological principles such that test data can be characterized as the realization of a stochastic event defined on a product space:

$$\Omega = \Omega_I \times \Omega_J \times \Omega_{\mathcal{S}},$$

where the orthogonal components,  $\Omega_I$ ,  $\Omega_J$ , and  $\Omega_{\mathcal{S}}$ , are the probability spaces for test items, test takers, and test settings (e.g., different test centers or online testing settings such as at home or workplace), respectively. Hence, modeling test data for concordance studies of the nature described in Zumbo (2021) minimally requires sampling assumptions of a hierarchical experiment (i.e., measurement process) about test items denoted  $I$ , test takers denoted  $J$ , and test settings denoted  $\mathcal{S}$  and the specification of a stochastic process that is supposed to have generated the data. We will limit our discussion to the three components. However, it should

be noted that the joint product space for these concordance studies can be expanded to include other spaces induced by raters or measurement occasions for repeat testers.

An example demonstrating the need for the new re-interpretation of the true score in a one-point-in-time cross-sectional assessment design for a widely used psychological instrument of causal attributional styles may help make the value re-interpretation of the CTT true score less abstract. Recall that this re-interpretation is an affordance of the measure-theoretic test theory characterization of CTT. Higgins et al. (1999) were interested in the psychological attribute “causal attributional style” assessed using the Attributional Style Questionnaire (ASQ), a self-report measure of the respondent’s attributional style. The ASQ, with twelve hypothetical events split evenly among positive and negative events from achievement and affiliation areas, asks participants to identify causes and rate each regarding their perceived locus, stability, and globality. Concerning the ASQ, if we focus, for example, on the negative life events, each rated according to the respondent’s perceived locus, stability, and globality, the settings may be characterized by the three causal dimensions denoted:

$$\mathcal{S}_{CausalDimensions} = (s_{Locus}, s_{Stability}, s_{Globality})$$

nested within the six negative life events, such as “you split up with your boyfriend/girlfriend”:

$$\mathcal{S}_{Events} = (s_{E1}, s_{E2}, \dots, s_{E6}).$$

Not surprisingly, this complex assessment design engendered debates surrounding attributional styles measured by the ASQ, which had centered on the questionnaire's psychometric properties (i.e., the item-level dimensionality). From my point of view, the debate was mainly about whether the complex structure reflected a measurement artifact (i.e., a nuisance method effect) or a more nuanced psychological theory of attributional style. Framed with the new re-interpretation of the true score of attribution or explanatory style, we concluded that despite assertions to the contrary in the research literature about causal attributional styles, we showed that it is not possible to eliminate the impact of situational characteristics on causal attributional style. Hence, as we concluded, one must account for the person in the situations relevant to the explanatory style, which is supported by the re-interpreted true score in the context of this essay and my explanation-focused view of validity.

**4.6.1.2. The Re-Interpretation of The True Score of CTT for a Repeated Assessment (Repeated Measures Assessment Design).** Rather than a cross-sectional measurement design, one could imagine that  $\mathcal{S}$  presents when the same test in the same ecological setting is administered to test taker  $j$ , in a repeated testing assessment design used to study the change or stability of the true score,  $T$ , for test taker  $j$ .

I first encountered a unique feature of the measure-theoretic test theory in the repeated testing assessment setting during a collaborative grant project with Brian Little and Donald Zimmerman at Carleton University in the late 1980s. Recall that test reliability is defined as a ratio of two components of variance, true score variance and error-score variance, with respect to a target population. It does not make much sense to discuss test reliability for an individual test taker because, in the conventional interpretation of true scores, a test taker’s true score is unchanging and immutable. Using the measure-theoretic test theory framework, we could define an individual's test reliability index using the re-interpretation of the true score and then, as suggested by Zumbo and Kroc (2019): (1) choose the manner in which to bound the error on measurement variation over time, (2) design the assessment experiment to actually measure the quantifier of interest, the estimand which in our case is the index of reliability based on the re-interpretation of the true score, and (3) the choose the estimator that meets the desired properties.

**4.6.2. Ecologically shaped or informed? Both concepts are important in understanding and**

---

### *creating sustainable assessment and testing systems*

I have chosen to use the term “ecologically shaped” throughout this essay; however, a reasonable alternative modifier is “ecologically informed” for observed and true scores. Both modifiers relate to the influence of ecological principles of the test context; see the list of relevant research in my program on this theme in the preceding sentence. However, they imply different levels of contextual (or ecological) engagement and application depending on the assessment setting and psychological attribute being assessed. One way to view the essential difference is that ecologically informed refers to uses, including inferences, claims, or decisions based on test or survey scores that take into account ecological knowledge and principles (for example, see Zumbo, 2017). It suggests that ecological considerations have been included in the thought process, potentially influencing outcomes. Ecologically shaped, conversely, implies that the ecological processes themselves have played a direct role in forming or influencing something. It suggests a more active and dynamic interaction with ecological forces, where ecological factors have molded the shape or structure of something over time (for example, see Zumbo et al., 2015).

In summary, being ecologically informed is about being knowledgeable and considerate of the context or ecology of testing or survey use, while being ecologically shaped indicates a direct and tangible influence of these ecological processes. As such, as an initial strategy, I tend to use “shaped” when referring to the observed or true scores and “informed” when referring to interpretation, judgments, test validation, and assessment use. Although practices for their use may be offered, both concepts are important in understanding and creating sustainable assessment and testing systems that align with the complex assessment setting described in the first section of this essay.

#### ***4.6.3. The origin story of the re-interpretation of CTT allowing for ecologically-shaped true scores***

A narrative description of the origins of the ecologically shaped observed and true scores emerging from an alternate re-interpretation of measure-theoretic test theory follows. It is evident from the sub-sections of this section of the essay that this re-interpretation of the central mathematical objects of CTT arose from simultaneous parallel lines of my research program that were influencing each other: (a) re-formulating mental test theory, including CTT and item response theory (IRT) as abstract mathematical models, using concepts in measure theory, probability theory, and functional analysis as appropriate, (b) development of an explanation-focused validity theory and validation methods, and (c) validation studies and assessment research more generally in international assessment and surveys, language testing, and quality of life and wellbeing, social indicators, and health and human development that influenced the first two lines of research (see, for example, Fox et al., 1997; Higgins et al., 1999; Hublely & Zumbo, 2013; Lane et al., 2009; Zumbo et al., 1993) that often required the derivation of variations of test theory models appropriate for the assessment setting by construction, not by assumption.

In 1995, Donald Zimmerman and I advanced our long-standing collaboration dating back to the late 1980s on the development of measure-theoretic test theory, a concept he initially outlined in his 1975 *Psychometrika* paper and earlier works from the mid-1960s. We were motivated by several goals. Two of the leading immediate goals were to (a) further understand the nuances and implications of the 1975 framework by continuing the development of an operator theory approach and (b) to put flesh on the bones and get a deeper understanding of a re-interpretation of the true score of CTT that had become part of our analysis and description, as described for example in the single-person reliability project with Zimmerman and Little described above, an affordance of results in Zimmerman (1975).

The most important developments in our program up to the year 2000 focused on the first immediate goal, as described in the paragraph above, and were reported in Zimmerman and Zumbo (2001), wherein we presented a model of tests and measurements that identified test scores with Hilbert space vectors and true and error components of scores with linear operators. This geometric formalism simplifies derivations in test theory and brings to light relations among concepts in probability, statistics, and measurement that are not otherwise apparent.

I was also motivated to derive a variant of CTT that permitted several cases of educational and psychological instruments and assessments that did not align with the Lord and Novick CTT model. The complex data structure did not match the hypothetical hierarchical experiment at the heart of CTT, with concern for experimental independence and uncorrelated errors commonly appended to the widely used variant of the CTT model. In addition, I grew concerned that the conventional Lord and Novick framework characterizes the measurement process as context-free. Lord and Novick's framework characterized the measurement process as *in vitro*, wherein any "extraneous" contextual, situational, and ecological variables were considered contaminants that must be stripped of the measurement process.

The CTT framework (reflected in, for example, Lord and Novick's axiomatization during what I refer to herein as the classical period of development) is not unreasonable if one considers that while individual differences have been central to human psychology since the early 20th century, the dominant individual differences model for mental testing that emerged is one in which, ironically, the individual effectively disappeared (Tolman, 1991). Danziger (1990) states, "[m]ental testing flourished because of an interest in individual differences, but this observation hides more than it reveals" (p. 107).

Indeed, the investigation of individual differences preceded the development of modern mental testing by many years. There were old interpretive practices of reading an individual's character with the help of bodily signs. These might be based on somatic indications, as in the classical doctrine of temperaments, or on facial characteristics, as in the relatively more recent versions of physiognomy. (p. 107)

As Tolman and Danziger note, this naïve model motivated the rapid uptake and development of psychometric methods that largely ignored the ensuing rich body of literature documenting the complexities of learning and human development by a primitive assumption about the homogeneity and linearity of data patterns that disguise what I have come to call the many ways of being human.

Although Zimmerman (1975) includes the essential elements to warrant this novel re-interpretation, Zimmerman and I wanted to learn more about the measure-theoretic model and highlight the advantages of an operator theory formalism that, among other things, would more naturally ground the re-interpretation of true scores and observed scores. Zimmerman and Zumbo (2001) highlighted that mathematical models based on linear operators also have been prominent in quantum mechanics. When first introduced into physics, Hilbert space concepts unified what had previously appeared to be two separate and distinct theories—Heisenberg's matrix mechanics and Schrödinger's wave mechanics. We noted that these theories turned out to be mathematically equivalent when reformulated in a Hilbert space setting by Von Neumann, Dirac, and others. Central to this line of thinking is that different mathematical models may be equally correct but allow for different interpretations that provide valuable insights into the phenomenon of interest.

Zimmerman and Zumbo's (2001) use of a geometric formalism, including linear operator and Hilbert space formalism, provided the level of abstractness that allowed us to investigate properties of CTT further, simplified derivations in test theory, and brought to light relations among concepts in probability, statistics, and measurement that are not otherwise apparent. In terms of the alternative re-interpretation of the true score of CTT, this formalism was meant to



provide a natural bridge to what Zimmerman and I imagined as a type of Everett interpretation or relative state formulation for measure-theoretic test theory in support of a re-interpretation of the true score and other objects central to measure-theoretic formulation of CTT. However, as you will see in the following paragraphs, the mathematic results in Zimmerman (1975) and Zimmerman and Zumbo (2001) were sufficient to warrant allowing a re-interpretation of the true score of CTT sufficient for our imagined purposes, and particularly as part of a coherent framework for my explanation-focused view of validity and validation, as well as assessment research more broadly without being drawn into the highly contested philosophical notion of a many-worlds interpretation of how the abstract mathematics of quantum mechanics relates to physical reality as we experience it on earth or elsewhere.

In summary, the re-interpretation of true scores rigorously defined in measure-theoretic test theory does not reflect the properties of test-takers (or survey respondents) but represents the properties of a test-taker or survey respondent defined by the assessment context or situation reflected in the measurement process. This measure-theoretic interpretation of the true score described by Donald Zimmerman and me in 2001 is reflected in the ecological (situational or contextual) item and test response model found in Zumbo et al. (2015).

Our program on a measure-theoretic test theory ended abruptly with Donald Zimmerman's death in December 2013. The loss of my mentor, longtime friend, and collaborator greatly delayed the introduction of the re-interpretation of true and observed scores in the psychometric "research literature." However, this re-interpretation of the true score informed many of our research studies collaboratively or separately. It is satisfying that our project achieved its immediate goals of a close study of the re-interpretation of true scores (and observed scores) in measure-theoretic test theory as contextualized, situated, ecologically informed observed score (and true score), as described in this essay.

The next two sub-sections describe the re-interpretation by first describing a summary of measure-theoretic test theory and, next, describing the interpretation of the true score based on its rigorous definition in measure-theoretic test theory. At the same time, Zumbo (2007b, 2009, 2015, 2017) traces how the re-interpretation of CTT as the ecologically shaped true and observed score (a) supports the explanation-focused view, (b) aligns with what Zumbo et al. (2015) refer to as the ecological model of item responding and test performance, (c) the ecology of item responding, as Zumbo and Gelin (2005) note, allows the researcher to focus on sociological, structural, community, and contextual variables, as well as psychological and cognitive factors, as explanatory sources of item responding, (d) third generation DIF (Zumbo, 2007b) as it relates to test validation, and (e) how a test taker's gender "... more properly should be considered a social construction, and gender differences on item performance are explained by contextual or situational variables (ecological variables, if you wish), such as institutionalized gender roles, classroom size, socioeconomic status, teaching practices, and parental styles" (Zumbo et al., 2015, p.139). Finally, and most importantly, Zumbo et al. (2015) provide an essential methodological focus that comes along with the re-interpretation of CTT in measure-theoretic test theory that

... it is important to keep in mind that we are adhering to the view that neither the test taker nor the cognitive processes in item responding are isolated in a vacuum. Instead, test takers bring their social and cultural present and history to test taking. We accept as our starting point the widely received view in the broader social sciences that human beings have evolved to acquire culture from birth and that the culture to which an individual is exposed, and the ecology of their lives, affects their basic psychology and cognition, including, in our case, item responding. In so doing, one can move to a contextualized form of explanation that works against a binary structure of variables that explain test performance. (p. 140)

Finally, drawing on the connection of DIF to the broader issue of measurement validity, the

ecologically shaped interpretation of the true and observed score and the ecological model of item responding and test performance further articulates what is meant by “context” in Zumbo’s (2009) view of validity as a contextualized and pragmatic explanation—that is, the multilayered ecology is the context.

#### ***4.6.4. Re-interpretation of the true score based on its rigorous definition in measure-theoretic test theory***

The purpose of this sub-section is to describe the interpretation of the score based on the rigorous definitions described above based on measure-theoretic test theory. Most importantly, Zimmerman (1975) defined true score as the conditional expectation of a test score when the conditioning is taken with respect to the test-taker. As described in the description of measure-theoretic test theory in the preceding sub-section, this mathematical framework allows for a re-interpretation of the true score (and hence the observed score) where one conditions on all possible outcomes of the measurement process  $X$  for a particular test-taker or survey respondent.

In short, the test theory models presented by Zimmerman (1975) and Zimmerman and Zumbo (2001) generalize classical test theory, allowing for a re-interpretation of true and observed scores reflecting in vivo (Zumbo, 2015), ecological item response and test performance (Zumbo et al., 2015), and as Zumbo et al. goes on to state, this re-interpretation aligns with the view that human beings have evolved to acquire culture from birth and that the culture to which an individual is exposed and the ecology of their lives affects their basic psychology and cognition, including, in our case, item responding and test performance, and finally this re-interpretation one can move to a contextualized form of explanation that works against a binary structure of variables that explain test performance.

Because the interpretation of the true score is situated or contextualized by the measurement process, my explanatory model of test score variation is likewise embedded within an ecological model of item responding that is situated within a pragmatic view of abductive explanation wherein one develops validity evidence for tests through abductive reasoning wherein, as I described in the previous section of this essay, the explanans are elements of my ecological model (Zumbo, 2007b), which may be involved in setting the initial conditions of my abductive method. As such, the item responses or test scores are the explanandum. In my explanation-focused view, my ecological model's constituent concepts and variables (Zumbo et al., 2015) are the explanans that, in short, explain the item responses or test scores (i.e., the explanandum).

As such, measure-theoretic test theory provides the basis for a coherent framework for my explanation-focused test validation and assessment research more generally. Raykov and Marcoulides (2011, pp. 119-121) provide a thorough and accessible description of the interpretation of true scores from a measure-theoretic vantage point.

#### ***4.6.5. Researchers don't always use the measure-theoretic test theory model, but when they do, they prefer the re-interpretation of the true score***

Remarkably, to my knowledge, Zimmerman’s (1975) landmark paper was largely ignored in the psychometric literature for the first decade and a half post-publication, as evidenced, for example, by it not even being mentioned by the eminent psychometrician Charles Lewis’ important paper reviewing developments in mental test theory (Lewis, 1986). The earliest exceptions to this are Steyer (1988, 1989) and Steyer and Schmitt (1990), who continued the development of CTT as related to the theory of conditional expectation based on principles in Zimmerman (1975) and its characterization in the context of confirmatory factor analysis (CFA).

Perhaps the most substantial research theme informed by Zimmerman’s (1975) model is the latent state-trait theory (LST), which, to my knowledge, was introduced by Steyer, Majeed,

~~Schwenkmezger, and Buehner (1989)~~ Steyer et al. (1989) within a CFA framework. These authors presented a generalization of classical test theory, LST, which explicitly considers the situation factor, introduced formal definitions of states and traits, and presented models in a CFA framework, allowing one to disentangle the effects of the trait and the effects of situations and/or interactions. What is most impressive about the LST developments in the latent variable and CFA approach is the rigor of mathematically well-defined true score variable definitions in line with their trait definitions and state factors. This level of rigor is not just a matter of mathematical virtue. However, it also justifies interpreting the latent variables and deciding whether or not it is, in fact, these variables that they are interested in for partitioning the state-trait variability.

LST has also been described and applied by Steyer et al. (1992), Steyer et al. (1999), and Geiser and Lockhart (2012). Moreover, similar to the developments in LST, two developments by Michael Eid stand out to me in this light. Eid (2000) developed a new model of confirmatory factor analysis (CFA) for multitrait-multimethod (MTMM) data sets that can be defined by only three assumptions in the measure-theoretic formulation of classical test theory. Furthermore, Eid (1996) describes the mathematical structure of several longitudinal confirmatory factor analysis models for polytomous item responses. Koch et al. (2014) describe a new longitudinal multilevel CFA-MTMM model for measurement designs with structurally different and interchangeable methods (Latent-State-Combination-Of-Methods model, LS-COM)- also see Koch et al. (2018).

I have only a passing knowledge of the extensive research literature going back over a century on personality psychology's aim to explain why people behave similarly or differently across time and contexts. This research literature also refers to this research purpose as the person-situation debate. Therefore, while learning about the details of LST for this essay, I only recently learned that Steyer and his collaborators share my view of the re-interpretation of true scores, which they describe in a latent variable setting involving multiple measurement occasions as persons-in-situations. To my knowledge, the first description of the latent variable CFA interpretation of persons-in-situations is described in Steyer et al. (1992).

Although measure theory is not a common language among non-mathematicians (Kroc & Zumbo, 2020), particularly among educational and psychological researchers and most assessment researchers, Zimmerman's (1975) rigorous characterization of CTT has become an important branch of a psychometric theory that has contributed to the development of LST by Steyer and, to my knowledge, many collaborators in his orbit and sphere of influence in the contemporary psychological research on the theory of states and traits, and psychometric traditions.

Separately from the research contributions in the development of LST, Raykov (1992, 1998a, 1998b) has an extensive research program building on the principles and results in Zimmerman (1975) to provide a deeper understanding of classical test theory in various psychometric settings, and also developed several univariate and multivariate models that emerged from an interaction between the classical test theory and the structural equation modeling approach. Raykov continues to be a prominent advocate of the rigor provided in what I call measure-theoretic test theory and Zimmerman's (1975) contribution. Raykov's body of research and substantial contributions to our understanding of psychometric methods are far too large to describe in detail; however, the following four stand out as building on or expanding the principles and results in Zimmerman (1975). First, Raykov (1992, 1999) significantly contributes to methods analyzing change over time. Second, Raykov (1998a, 1998b, 2001) makes significant developments in psychometric theory concerning test reliability and standard error. Third, Raykov and Marcoulides (2016) describe the relationship between classical test and item response theories. Fourth, Raykov and Marcoulides (2011) is the first English text on

psychometric theory that devotes considerable attention to the principles and critical results of Zimmerman (1975).

#### **4.7. How Perspectival Realism and Pragmatic Undercurrents of Conditionalized Realism Inform My Explanation-Focused Validity Theory and Assessment Research**

Let me reiterate that from my description of my view of the philosophy of scientific realism in the third section of this essay, my view of scientific realism is closest to Giere's (2006) perspectival realism with pragmatic undercurrents of Schaffner's (1993) conditionalized realism. As reflected in Zumbo (2009) and Stone and Zumbo (2016), my views continue to reflect a substantial pragmatic component; Schaffner's (1993) "conditionalized realism" shaped my earliest theoretical developments in validity theory and continues to do so. Although there are recognizable differences between them, I do not find the concepts of conditionalized realism wholly incompatible with the perspectival view.

As such, I do not embrace a strong anti-realist stance in my assessment research and theorizing. Nevertheless, I also reject a wholly committed realism. In this way, I agree with Schaffner that we do not have any direct intuitive experience of the certitude of scientific hypotheses or theories. Furthermore, importantly for validity theory, regarding entity realism about psychological traits and latent, hypothetical, intervening, or latent class variables, Green's (2015) description reflects mine well: "I am fairly realist about some scientific objects (e.g., trees, mountains, stars) and I am fairly instrumentalist (anti-realist) about others (e.g., the implicit memory system, the openness-to-experience personality trait, dissociative identity disorder" (p. 212). Green goes on to state that:

Finally, I took a short excursion into philosophy of science, trying to explain how antirealism is not, in the main, antiscience but, rather, an effort to come to terms with the history of science as it has actually proceeded over the past several centuries. One need not conclude that there are no "real objects out there" in order to see the power of the antirealist narrative. (p. 212)

A few remarks may help sketch out the form of my view. I tend toward a perspectival realist view that argues that the specific "viewpoints" within which scientists must work do not prevent them from discovering objective reality features. Giere describes his characterization of much scientific knowledge as "perspectival realism."

I will explain perspectival realism in the setting of assessment research as involving two parts. First, claims about the psychometric properties of an assessment, such as the item characteristics, the dimensionality of the item responses, or differential item functioning (DIF) generated by the scientific practice of validation research are claims about the world and not, for example, claims about beliefs about the world. If you wish, making claims about the world rather than beliefs about the world is the realism part. Second, these claims are not absolute or without conditions or limitations; they are thus conditional, which is the perspectival part of perspectival realism. It is noteworthy that the kind of conditionality considered in perspectival realism needs to go beyond the widely held case that claims about the psychometric properties of an assessment instrument are limited to the current body of evidence about an attribute or construct of interest because that is a low bar for conditionality. That is, the perspectival part of perspectival realism has to add more to scientific practice and discourse than the widely accepted conditionality that our knowledge claims about the material world are limited by (conditional on) our current body of evidence, which few scientists would question. In a sense, the conditionality has to add more value than conditioning on something few would question.

Adapting Giere's description, these claims about the psychometric properties, such as the DIF of the item responses, are not absolute but relative to humanly constructed concepts or "conceptual schemes" such as the DIF method (e.g., logistic regression or Stout's SIBTEST) and the grouping variables which from the vantage point of the many ways of being human are

---

not natural kinds which I discuss in more detail in section five of this essay.

A more nuanced example of conditional claims relative to different conceptual schemes is described in section four of this essay, where the redefinitions and results of the true score from (the conceptual scheme of) measure theory and functional analysis (Zimmerman & Zumbo, 2001) inspired the analytical methods of my explanation-focused view of validity. We also see that the definitions of true scores relative to the different conceptual schemes free up the interpretation so that users of test theory are not forced to invoke a trait view of the true score. The main point of this example is that claims about the properties of true scores are conditional to the psychometric conceptual scheme. The reader should note that I am taking some expository liberties equating a conceptual scheme with a formal system of a given mathematical theory.

To conclude the description of Giere's perspectival realism, as he reminds us, it is notable that the perspectivism involved is not global but confined to scientific knowledge, so it is a scientific perspectivism. In addition, it is important to note that the presupposed conceptual scheme is the property of a scientific community. Therefore, for example, I would argue that Zimmerman and Zumbo's (2021) geometry of probability, statistics, and test theory would be said to provide a measure-theoretic test theory perspective on observed item or test response data that arise from the measurement processes. Of course, this perspective could be contrasted with an item response theory (IRT) theoretical perspective on the observed item response data, noting that as Kroc and Zumbo (2020), in terms of their mathematical structures, in no way is the classical test model equivalent, or even necessarily comparable, to the IRT measurement model.

Regarding how my particular realist stance informs my explanation-focused leanings, the science of assessment and testing I am advancing in this essay does not look simply for theories compatible with the attribute we wish to measure but are true, explanatory, and fecund. As such, the kinds of explanatory theories I imagine must be plausible (i.e., consistent with the largest possible background of accepted beliefs and reflect the many ways of being human), empirically testable, and provide models of the response processes that support claims of the validity of the inferences, claims, and uses of test and assessment scores, and which at the same time explain the variation in item or test scores from my ecological model of item or test responding.

Within a perspectival tradition, I suggest that while creating explanations (explanatory theories, if you wish), assessment scientists create perspectives, in Giere's sense, describing and conceiving aspects of the assessment data that may include item responses, test scores, information about the test taker, and what Zumbo et al. (2023) describe as sensor data (e.g., eye tracking or response latency) that arise from the testing encounter in computer-based measurement processes. It is important to note that with what Zumbo (2023a) describes as putting psychology back in psychometrics, the focus is on the "encounter" of a person and an item (or task); this is defined as the interaction of person and item in the measure-theoretic view of test theory (Zimmerman & Zumbo, 2001).

Finally, in line with the perspectival approach, Zumbo (2007a) highlighted and adapted for his explanation-focused view a point by Suppes (1969) and Woodward (1989), explanatory models, including psychometric models, are typically compared not directly with experimental data but with models of data. A long tradition and many different statistical techniques may be used in deciding when the observed agreement is sufficient to infer a general fit between the model and the assessment data arising from the measurement process.



## **5. DESCRIPTION OF MY EXPLANATION-FOCUSED VALIDITY**

With the preliminaries of explanation-focused validity behind us in section 4, this section aims to describe the current version of my explanation-focused view of assessment research and test validity and how it has developed into a coherent research framework for test validity and assessment research. More generally, my explanation-focused test validity and assessment research is embedded within an ecological model of item responding and test performance, placing a centrality on test consequences and values and what I refer to as the many ways of being human (Zumbo, 2018a). This section of the essay integrates and builds upon the description of explanation-focused validity theory and aligned validation practices that began with the 2005 Messick Award address at the joint annual meeting of the International Language Testing Association (ILTA) and the Language Testing Research Colloquium (LTRC) (Zumbo, 2005, 2007a, 2009, 2017, 2021, 2023a, 2023b).

### **5.1. Explanatory Considerations in Test Validation and Assessment Research**

At the core of my view of validity is that we should aim to build a science of educational and psychological test validation; a good science does not merely describe and predict phenomena but must explain them.

A concise statement of my explanation-focused view is: “[v]alidity is a matter of inference and the weighing of evidence; however, in my view, explanatory considerations guide our inferences” (Zumbo, 2009, p. 69). Zumbo (2009) described how his explanation-focused view builds upon Messick’s (1989) description of test validity involving

“... an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment.” (p. 13)

a step further to argue that conventional validation practices (e.g., reliability coefficients, validity coefficients) are descriptive rather than explanatory and that validity should, in addition, provide a richer explanation for observed test score variation.

In this sub-section of the essay, I will briefly describe the basic idea underlying it, how I conceive of explanation as abductive and inference to the best explanation, and insights into it by reconsidering it alongside five conceptions of validity described in the historical analysis that invoke some form of explanation.

### **5.2. Basic Ideas Underlying My Explanation-Focused Validity: Bridging the Inferential Gap, Abductive Methods, Inference to the Best Explanation, and Explanatory Coherence**

My explanatory approach's basic idea is that understanding the item or task score variation would go a long way toward bridging the inferential gap between test scores (or even latent variable scores) and educational or psychological attributes we purport to measure. According to this view, validity, per se, is not established until one has an explanatory model of the variation in item responses, the scale scores, or both, and the variables mediating, moderating, and otherwise affecting the response outcome.

This expectation is a tall hurdle indeed. Overlooking the importance of explanation in our definition of validity and hence reflecting it in our validation practices, we have, as a discipline, focused overly heavily on the validation process. As a result, we have lost our way. This statement about the importance of explanation is not intended to suggest that the activities of the validation process, such as correlations with a criterion or a convergent measure, dimensionality assessment, item response modeling, or differential item or test functioning, are irrelevant or should be stopped. Quite to the contrary, the activities of the validation process must serve the definition of validity. I aim to re-focus our attention on why we are conducting all of these psychometric analyses: to support our claim of the validity of our inferences or decisions from a given measure.

---

In my view, validity is a matter of inference and the weighing of evidence; however, in my view, explanatory considerations guide our inferences. Explanation acts as a regulative ideal; validity is the explanation for the test score variation, and validation is the process of developing and testing the explanation.

As the basis of measurement science, explanatory considerations guide our inferences or claims resulting from reporting or using scores, or both (Zumbo, 2005, 2007a, 2007b, 2009, 2017). Thus, in my view, construct validation should seek an explanation of the items' or test score variation or the variation in the outcome of test use, for example, using the test scores to classify or decide a test-taker's standing according to a standard-setting exercise. Although I will unpack this further in a subsequent section of this essay, it is noteworthy that I do not take as a first principle that the hypothetical construct as per Cronbach and Meehl (1955) or as per Borsboom et al. (2004) the latent variable as a conceptual mapping of the empirical phenomenon as a conceptual mapping of the empirical phenomenon explains the test score variation.

I devoted attention to describing my definition of test validity because I hold as essential that if one wants to advance the theorizing and practice of measurement, I believe one needs to articulate what they mean by “validity” to go hand-in-hand with the validation process (Shear & Zumbo, 2014; Zumbo, 2007a, 2009;).

Notably, I consider test validity centrally involves making inferences of an explanatory nature (Zumbo, 2007a, 2009); however, depending on the type of assessment, such as self-report ratings, task performance, knowledge, or achievement items on an educational assessment that are scored correct/incorrect or for partial knowledge, for example, and the extent and richness of the background knowledge, there are somewhat different patterns of abductive inference. Of course, when our initial understanding of the psychological attribute is thin, our most convincing explanation may amount to mere conjecture.

Zumbo (2005, 2007a, 2009) described an explanation-focused approach to test validity in which test validation centrally involves making inferences of an explanatory nature, highlighting inference to the best explanation (IBE). This reliance on explanation and IBE was presented contra the dominant mode of construct validation framed as hypothetico-deductive empirical tests in line with Cronbach and Meehl and those scholars who advocated that view. My view of test validity is also meant to guide our assessment research and reflects my perspective that validity: “[e]xplanation acts as a regulative ideal; validity is the explanation for the test score variation, and validation is the process of developing and testing the explanation” (Zumbo, 2009, p. 69).

I cannot stress this enough that in terms of the process of validation as opposed to test validity itself, the statistical methods, as well as the psychological and more qualitative methods of psychometrics, work to establish and support the inference to the best explanation (IBE)— i.e., validity itself; so that validity is the explanation, whereas the process of validation involves the myriad methods of psychometrics to establish and support that explanation. Interestingly, it is notable that IBE essentially combines the justificatory and explanatory sorts of arguments; first, we formulate an explanation, then a justificatory argument to convince us it is indeed the best possible explanation.

In line with the perspectival and pragmatic undercurrents of conditionalized realism described in the previous section of this essay, IBE informs my explanation-focused validity theory and assessment research; however, as highlighted by several philosophers of science, except for Thagard's (1992) description of IBE as centrally concerned with establishing explanatory coherence, typically judgments of the best explanation primarily provide grounds for acceptance of the explanatory model or theory. My most recent developments explicitly incorporate Thagard's explanatory coherence (1989) into the description of the higher-order

integrative cognitive process model, involving every day (and highly technically evolved) notions like concept formation and the detection, identification, and generalization of regularities in data, whether numerical or textual. From this, after a balance of possible competing views and contrastive data, comes understanding and explanation (Zumbo, 2009, pp. 69-70). Haig (in press) describes a modern variation of Thagard's coherent explanation that, in the end, future research may have several demonstrable advantages over the comparatively rudimentary strategy described here.

### 5.3. Exploratory Factor Analysis, Latent Variable Regression Models, and the Pratt Index for Variable-Ordering as Examples of Explanation-Focused Validation Methods

#### 5.3.1. Exploratory factor analysis, theory generation, and scientific method

Haig (2005a, 2005b, 2009, 2018, in press) makes a compelling case for factorial theories and factor analysis, particularly exploratory factor analysis (EFA), as an abductive method of theory generation that fits well with my explanation-focused view of validation. He states that the factorial theories reflected in the findings of EFA are essentially dispositional and that invoking a form of existential abduction provides us with an essentially dispositional characterization of the latent entities EFA postulates. He cautions that on their own, these dispositional explanations have limited, yet still valuable, explanatory import.

Recent developments in factor analysis methodology by Wu et al. (2014) that build on Zumbo's (2007a) introduction of variable ordering methods, referred to as Pratt methods, will likely be valuable in using EFA for explanatory validation purposes. In particular, what Wu et al. refer to as horizontal interpretation will aid in disentangling the effect of the latent factors on item responses by decomposing the factor loadings and communalities across the latent factors for each item one at a time. Essentially, the horizontal interpretation considers factors as the underlying causes that explain the common variation in item responses (or sub-scale variation).

#### 5.3.2. Partitioning the explanatory variation using a novel latent variable regression with a Pratt index

Zumbo (2007a) describes the Pratt indices and how they can be applied to a latent variable regression model, a variation of the classic multiple-indicators multiple causes model. Using data from the 20-item version of the original Center for Epidemiologic Studies Depression (CES-D) self-report measure, each item has a 4-point response format. I demonstrated how a researcher interested in the working hypothesis of the postulated explanatory role of a respondent's age and gender on test performance, i.e., the CES-D overall scale score for its 20 items, based on extensive prior research reported in the scientific literature.

In order to describe the latent variable regression model, we can first describe the typical confirmatory factor analysis (CFA) model, in which the score obtained on each item is considered a linear function of a latent variable and a stochastic error term. The linear relationship may be represented in matrix notation, assuming  $p$  items and one latent variable as

$$y = \Lambda \eta + \varepsilon, \quad (3)$$

where  $y$  is a  $(p \times 1)$  column vector of continuous scores for person  $j$  on the  $p$  items,  $\Lambda$  is a  $(p \times 1)$  column vector of loadings (i.e., regression coefficients) of the  $p$  items on the latent variable,  $\eta$  is the latent variable score for person  $j$ , and  $\varepsilon$  is  $(p \times 1)$  column vector of measurement residuals. However, the latent variable regression model for the CES-D includes ordered categorical item response data; therefore, for item  $j$  with response categories  $c = 0, 1, 2, \dots, C-1$ , define the latent variable  $y^*$  such that

$$y_j = c \quad \text{if} \quad \tau_c < y_j^* < \tau_{c+1},$$

where  $\tau_c, \tau_{c+1}$  denote the latent thresholds on the underlying latent continuum, which are typically found to be spaced at non-equal intervals and satisfy the constraint  $-\infty = \tau_0 < \tau_1 < \dots < \tau_{c-1} < \tau_c = \infty$ .

To write a general model allowing for predictors of the observed (manifest) and latent variables, one extends equation (1) with a new matrix that contains the predictors  $x$

$$y^* = \Lambda z + Bx + u, \quad \text{where} \quad (4)$$

$$z = Dw + \delta,$$

$u$  is an error term representing a specific factor and measurement error and  $y^*$  is an unobserved continuous variable underlying the observed ordinal variable denoted  $y$ ,  $z$  is a vector of latent variables,  $w$  is a vector of fixed predictors (also called covariates),  $D$  is a matrix of regression coefficients and  $\delta$  is a vector of error terms which is distributed  $N(0, I)$ . Finally, as described by Zumbo (2007a, p. 65-71), using the Pratt index, 61.5% of the explained variation (i.e., the R-squared) in the observed CES-D total scale score is attributable to the age of the respondents, and the remainder of the explained variation reflects the gender difference; this makes age the more important of the two predictors.

#### 5.4. The Ecological Model of Item Responding and Subtest or Test Performance: A Conceptual Model

Zumbo et al.'s (2015) purpose for introducing the ecological model of item responding was to move beyond the simple explanatory ideas embodied in the widely used psychometric models like item response theory or factor analysis that a single unitary latent variable is the sole explanatory variable that explains the pattern in item responses (Goldstein, 1980; Goldstein & Wood, 1989).

Instead, the aim was to foster psychosocial theorizing about item response processes contributing to an emerging paradigm shift in measurement, survey design, and testing wherein one embraces the diversity of test takers (their histories, communities, cultures, and life experiences) and leverages developments in data science, computation, technology, and psychosocial theories to do principled assessment reflecting the many ways of being human in our contemporary world and to do it in a valid and effective way. This new form of differential item and task analysis is a critical component of my new psychometric paradigm that has laid the groundwork to expand the evidential basis for test validation by providing a richer explanation of the processes of responding to tests, promoting richer psychometric theory-building.

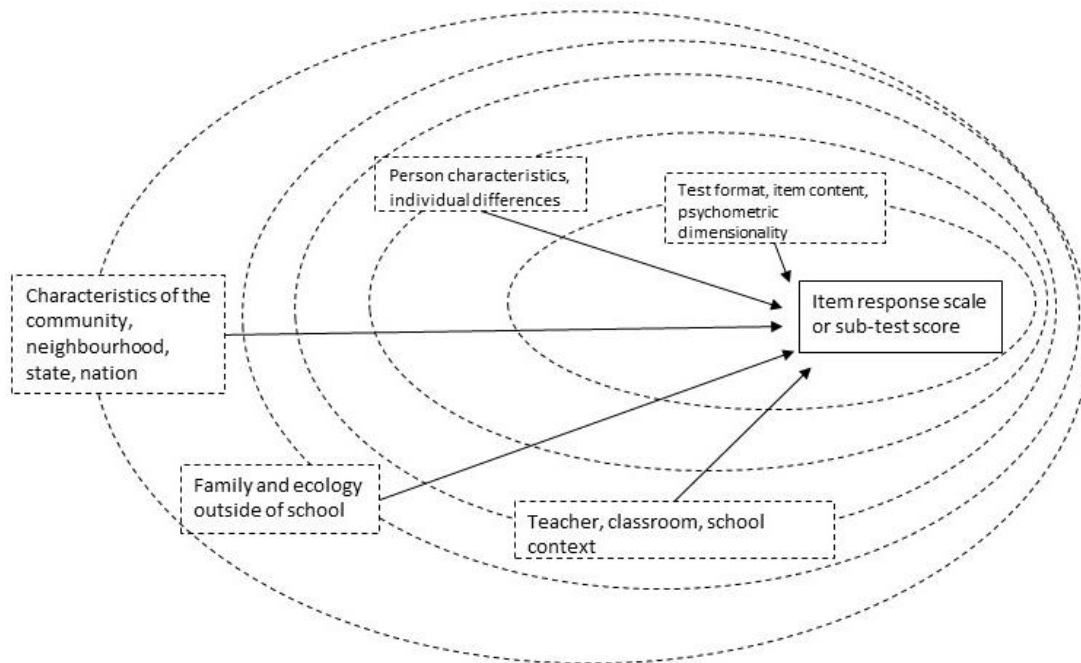
Allow me to make a critical sidebar remark before describing the ecological model. I do not wish to be interpreted as suggesting that I am the first or sole psychometrician to take aim at conventional assessment research and validation practices and offer an alternative item response model, which is far from it. For instance, focusing on the statistical models widely used in assessment research, Goldstein (1994) raises several important critical points. For example, he addresses a serious question about traditional exploratory factor analytic techniques when one unknowingly has more than one subpopulation under study when the test takers' item responses are related to their gender or levels of education. It is possible that the resultant omnibus latent variable model fits well in one subgroup but not in the other subgroup(s). Also, Goldstein & Wood (1989) note that one or more of the latent variables that emerge from an analysis of a heterogeneous population may be explained by such factors as, in

our example, gender or levels of education. However, Goldstein (1995) offers an elegant solution where differences due to, for example, gender or levels of education could readily be incorporated into multivariate item response models used to provide fully efficient estimates.

Let us return to my ecological model of item and test performance. Figure 1 is a graphical portrayal of an instantiation of my ecological model characterized by five concentric ovals representing potential explanatory sources of variation for item, test, or scale scores adapted from Zumbo et al. (2015), who consider an example of large-scale educational testing. Depending on the in vivo assessment setting, different explanatory sources may be described as concentric ovals. Likewise, the outcome variable of the explanatory variables may be (i) responses to an item or task, (ii) a sub-scale score from the assessment or a dimension of a multidimensional measure, or (iii) an overall test score. Moreover, depending on the statistical psychometric model, the sub-scale, dimension, or overall score could be modeled as a latent outcome variable.

In the example above, the sources of variation depicted in the concentric ovals represent (a) test format, item content, and psychometric dimensionality; (b) person characteristics and typical individual differences variables such as cognition; (c) teacher, classroom, and school context; (d) the family and ecology outside of the school; and finally (e) characteristics of the community, neighborhood, state, and nation. Conventional first and second-generation DIF practices have focused on the first oval with some modest attempts at the second oval as a source for DIF explanation. In contrast, the emerging paradigm from my research program takes an ecological modeling approach informed by Bronfenbrenner’s ecological systems theory (e.g., Bronfenbrenner, 1979, 1994).

**Figure 1.** *An instantiation of the ecological model of item or test performance adapted from Zumbo et al. (2015).*



Wallin (2007) describes a potentially fruitful explanatory view of how environmental considerations (which I describe as ecological) help us explain a psychological process's form or function as follows. Adapting Wallins’ description, an ecological explanation thus allows one to frame the explanatory considerations about the function of a process (e.g., the mental processes involved) in relation to the ecology in which the process is active (depending on the



micro, meso, or macro components of Zumbo et al.'s model), and to the adaptive value of the function in the environment under consideration, with particular attention to cultural adaptation (p. 164). Wallin states, "An ecological explanation explains the design of a psychological process by referring to the adaptive value of the design given a particular environment, and a particular function" (p. 163), thus moving the ecological model of item and test performance substantially closer to explanatory coherence.

### **5.5. An Ecologically Informed, In Vivo View Describes the Enabling Conditions for the Abductive Explanation**

The ecologically informed in vivo view of validation practices describes the enabling conditions for the abductive explanation for variation in test performance (Stone & Zumbo, 2016; Zumbo, 2007a, 2009). As such, the study of response processes is guided by a contextualized pragmatic form of abductive explanation. In terms of the process of validation (as opposed to validity, itself), the methods described herein work to establish and support the inference to the best explanation – i.e., validity itself; so that validity is the contextualized explanation via the variables offered in the ecological model, whereas the process of validation involves the myriad methods of psychometric and statistical modeling (Zumbo, 2007).

Zumbo's abductive approach to validation seeks the enabling conditions via the ecological model through which a claim about a person's ability from test performance makes sense (Stone & Zumbo, 2016; Zumbo, 2007a, 2009). In contrast to inductive or deductive reasoning, abductive reasoning neither construes the meaning of the scores purely from empirical evidence nor assumes the meaning of the test to explain the score. Instead, abductive reasoning seeks the enabling conditions under which the score makes sense. The reader unfamiliar with these forms of reasoning is encouraged to consider Haig's (2019) assessment of three major theories of the scientific method: hypothetico-deductive method, inductive method, and inference to the best explanation. He describes a broad abductive theory of scientific method that has particular relevance for education and psychological assessment research and validation practices.

In short, abductive reasoning and the inference to the best explanation aim to explain why people behave similarly or differently across, for example, time and contexts – an alternative expression I have used is how well a test or assessment travels across time and place.

Appropriate modeling strategies must include various aspects of the ecology framework within a single set of analyses. The multilayer nature of item-responding ecology fits well with multilevel modeling via mixture models. Lower-level observations are nested within a higher-level factor within a hierarchical system. This nesting nature of observations is likely to produce some degree of similarity among the observations nested within the same unit. Thus, these observations are not entirely independent from each other. This nesting does not imply that variables drawn from personal characteristics and family ecology should always be modeled in a multilevel regression model at different levels. The level of the variable being measured, the structure of the data, and the theory to be tested must be considered when deciding upon the structure of a multilevel model.

A natural question arises of how the ecological model's contextual factors could affect item responses by mediating the cognitive processes normally assumed to generate item responses. Even glancing at [Figure 1](#) may raise the question of how any proximal ecological variables, such as neighborhood characteristics, in my model impact the cognitive processes, writ large, involved in the item response when they are, in essence, so far away from the item, subscale, or test performance. In response, I question the exclusive emphasis on individual test-taker characteristics such as cognitive factors and argue that greater attention must be paid to basic social conditions if this new ecological paradigm is to have its maximum effect in the time ahead. There are two reasons for this claim. First, I argue that test-taker cognitive factors must be contextualized by examining what puts people at risk of performing poorly (or, equivalently,

performing well) on educational achievement tests if we are to craft educational interventions that improve learning opportunities for all, per my view of the many ways of being human. Second, I argue that ecological factors such as socioeconomic status or access to support that characterize, for example, the student's family or ecology outside of school or even further proximally as a neighborhood characteristic *may be* more fundamental than the personal characteristics for specific educational assessments tracing learning and educational progress because they embody access to important resources, affect multiple intermediary learning processes and outcomes through multiple mechanisms. Without careful attention to these possibilities of the import of proximal ecological variables, we risk overlooking sources of hidden invalidity.

In short, one of the central features of this ecological framework is that it explicitly illustrates the complexity of the ecology of the item response. This ecological framework is proposed to motivate a focus on contextual factors and to guide the development of contextual models to explain item responses via these enabling conditions guiding the abductive explanation. Without a conceptual framework organizing various aspects of the ecology of item responding, it is difficult to systematically study the sources of item response or test performance variability.

## **5.6. Test Validity in The Context of Concomitant Changes in the Value-Free Ideal in The Philosophy of Science**

Taking a lesson from the confusion and misunderstandings of Messick's description of the role of values and test consequences in assessment research and validation, I have devoted a subsection of this essay to discussing the value-free ideal and test consequences. I will focus my remarks on what I describe as concomitant changes in test validity, including my explanation-focused theory and changes in the value-free ideal in philosophies of science.

### **5.6.1. Value-laden stance that guides the question of epistemic integrity**

Nearly concomitant with Messick's (1980, 1989, 2000) theoretical developments in a validity theory that viewed values and consequences as an integral part of construct validity and the validation process as they contribute to the soundness of score meaning, developments in the philosophies of science were beginning to consider a value-laden stance that guides epistemic integrity. This philosophical tradition focused on epistemic integrity in the epistemology of science; the suggestion was that there are two distinct notions of research integrity in use—an epistemic notion, which focuses on the reliability of the research results, and a moral notion, which concerns the moral acceptability of research practices.

### **5.6.2. Brief description of philosophy's response to the value-free ideal**

Douglas (2016) reminds us of the philosophical doctrine dating back to the mid-1700s that an evaluative statement cannot be derived from purely factual premises, implying no logical connection between facts and values (sometimes referred to as Hume's Law) was, in good part, the inspiration for the value-free ideal: the idea that science is (or at least ought to be) free from values. Advocates of this value-free ideal make a case for the clear separation of (a) fact and value, (b) the descriptive and the normative, and (c) science and a set of opinions or beliefs of a group or an individual. Furthermore, these advocates for the value-free ideal acknowledge that day-to-day scientific practice is not always wholly free from values, but they insist it should be.

In contrast to the value-free ideal, advocates of the position that one should pay attention to values highlight the importance of value to make arguments explicit. In scientific practice, the expectations generated by a scientific idea and the actual observations relevant to those expectations form what is widely called a scientific argument. In scientific practice, one should be able to articulate the premises of our arguments with an aim for others in our scientific community to inspect and assess our reasoning. This kind of transparency is essential to any

self-correcting epistemic community, including identifying the nature of disagreement among scientific arguments. As such, advocates of this position argue that an awareness of values is also necessary to establish that our arguments are sound. The aim is for scientists to become aware that their scientific arguments rely on normative premises, forcing them to subject them to critical scrutiny and to show that the premises are true and, in the end, the arguments sound. A final reason for acknowledging the role of values in science is most evident at the intersection of science and policy-making. Beyond acknowledging the role of values in the day-to-day practice of science, it is essential to acknowledge how values inform policy decisions to empower the stakeholders fully.

ChoGlueck (2018) states in the opening of their paper that "... increasingly, philosophers have rejected value-free ideals of science and turned their attention to examining values in concrete cases and developing alternative norms for legitimate/illegitimate influences (see Hicks 2014)". They succinctly describe the current state of affairs in the philosophies of science.

The value-free ideal of science narrows the role for social, ethical, and political values—taken to be distinct from scientific, epistemic, and cognitive values—in scientific reasoning and practice (Douglas 2009; Elliott 2011). Defenders of this value-freedom accept the legitimacy of social, ethical, and political values only in the early and late stages of science, such as with funding and technological applications. The ideal proscribes the use of these purportedly nonscientific values within the so-called internal core of scientific reasoning, especially in evaluating evidential support for a hypothesis (i.e., theory choice). (ChoGlueck, 2018, p. 705)

Holman and Wilholt (2022) make a case that, given the widespread acceptance among philosophers writing about values in science that "... values necessarily play a role in core areas of scientific inquiry, attention should now be turned from debating the value-free ideal to delineating legitimate from illegitimate influences of values in science" (p. 211).

We will return below to what the field of assessment and testing can gain from philosophy's recent social turn.

### ***5.6.3. Brief review of assessment and testing's response to the value-free ideal***

One can see threads of this concern over value-free science interwoven in the various debates in the assessment and testing literature of the late 1980s and 1990s when the inclusion of broader social consequences and the inclusion of negative unintended as well as positive intended consequences by Messick (e.g., 1980, 1989) led to objections by several assessment researchers (see, for example, Green, 1990; Mehrens, 1997; Popham, 1997; Wiley, 1991).

In support of Messick's research program, assessment researchers like Hubley and Zumbo (1996) and Shepard (1997) argued that awareness of values is necessary. Shepard highlights that consequences have already been accepted as part of test validity for several decades and are a central part of the evaluation of test use. Kane (2006, p. 54) has recently noted that there is, in fact, nothing new about giving attention to consequences in investigations of validity. What is relatively new is the salience of the topic and the breadth of the reach that is no longer limited to immediate intended outcomes (e.g., test takers who access test preparation materials perform better on the test). We will return shortly to this matter of the salience and breadth of the reach when we discuss caveats for considering social aspects of assessment and the consequences of testing.

In the last 30 years, several researchers in validity theory (e.g., Addey et al., 2020; Hubley & Zumbo, 1996, 2011; Kane, 2016; Markus, 1998; Messick, 1998; Zumbo, 1998, 2017) have been pursuing this research agenda similar the one described by Holman and Wilholt (2022) that turns its attention from debating the value-free ideal to delineating legitimate from illegitimate influences of values in science. These assessment researchers implicitly or explicitly consider these questions from different theoretical orientations inspired, in large part, by Messick's research program. Given the central role of the question of the role of values in the interpretation

by some assessment researchers of the centrality of the concept of test consequences and Hubley and Zumbo's (2011) description of social and personal consequences and side effects, this essay is a continuation of that research legacy.

From our point of view described in this essay, arriving at a claim about social and personal consequences and side effects involves conceptualizing it as evidence/data-based policy-making that is essentially tied to test validity and establishing an evidential trail that supports that the proposed social and personal consequences and side effects are not unreasonable and are reproducible and generalizable, akin to more widely accepted day-to-day scientific practices. Designing and implementing assessment research according to best practices matters for the sake of the test's integrity, reliability, and validity and as necessary evidence in defending the test interpretation and use if challenged by critics and in a test review. Therefore, a method for test validation and accompanying considerations of social and personal consequences and side effects is not solely about statistical considerations; the statistical considerations should shine a light on the right questions and help resolve them.

The evidential trail is critical to the whole process because it is widely recognized that there is an element of judgment in all assessment and test validity research that is arbitrary in the sense that a range of legitimate choices could be made- for example, the various frameworks to test validation we described near the start of this essay.

#### ***5.6.4. Building on Messick's legacy of the role of values and consequences and recent developments in the philosophy of science***

The following remarks draw on a series of invited addresses (Zumbo, 2016, 2016b, 2018a) to assessment practitioners. Zumbo presents a contemporary perspective on test validation and a new view of measurement science that (like Messick before him) recognizes that values necessarily play a role in core areas of test design, delivery, and validity, taking the first steps in delineating legitimate from illegitimate influences of values in science. As highlighted by Messick, the challenge to future developments in assessment research and studies of test validity must reconcile our disciplinary history of naïve objectivity, the value-free ideal, and the inherent value-ladenness at the core of test validation.

Building on the philosophical and methodological writings of Douglas (2000, 2003, 2004) and others, it is evident from the description above of validation research that declaring someone has met a language assessment standard for immigration purposes (or some such claim) based on test results is a kind of type of claim about a phenomenon of interest to science whose definitions rely on a normative standard. Normative statements make claims about how things should or ought to be, how to value them, which things are good or bad, and which actions are right or wrong. Empirical generalizations about them thus present a special kind of value-ladenness. Philosophers of science have already reconciled values with objectivity in several ways. None of the existing proposals are suitable for the claims made in testing and assessment – what Zumbo (2016b) described as a blending of normative and empirical claims in his address. He argued that empirical claims from test performance have such a “blended” structure. Some say that these “blended” claims should be eliminated from science. Our position is that we should not seek to eliminate them from the science of measurement and testing. Instead, we need to develop principles for their legitimate use. We articulate a conception of objectivity for our science of measurement and testing that embraces these “blended” claims. Douglas (2004) gives us some direction on this front, but it is just a start.

In an important sense, this essay is built on an initial articulation of these new rules and strategies to secure procedural objectivity for measurement and testing. Find/discover the hidden value propositions in the tests and measures. This discovery of hidden value propositions needs to be systematic and documented as part of the process and needs, in part, to focus on disagreements about the empirical claims from the test. Check if value

presuppositions are invariant or robust to these disagreements, and if not, conduct an inclusive deliberation involving test performance data. Kane's (2012, 2013, 2016) approach to validation and Addey et al. (2020) are particularly well suited for this purpose.

### 5.7. Explicit Synthesis of Explanation-Focused and Argument-Based Approaches to Test Validation

Synthesizing the explanation-focused and argument-based approaches aims to close the gap between validity theory and the practice of validation such that test-score interpretations and uses are supported by appropriate evidence and reduce the chances of hidden invalidities. Zumbo (2023b) describes a test validity framework depicted in [Figure 2](#), synthesizing explanation-focused validation and argument-based approaches that incorporate features of his earlier work (Hubley & Zumbo, 2011, 2013; Zumbo, 2017, 2023a; Zumbo & Shear, 2011) which build on earlier work by Messick (1995, 1998, 2000) and reflect the principles of argument-based validation practices. A diagrammatic representation of test validation aims to depict the complex evidential bases of test validation, their interrelation, and their foundation on values.

As stated by Zumbo (2023b):

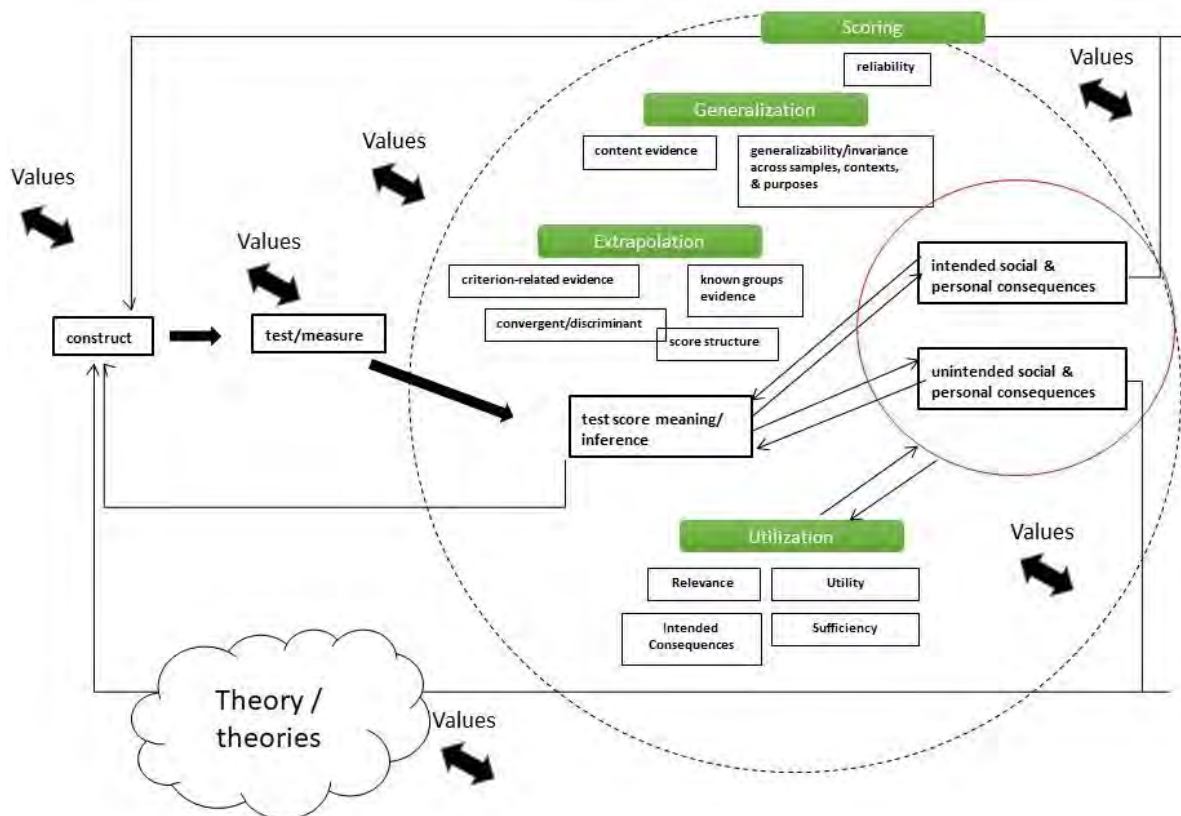
Since the publication of Messick's groundbreaking review of validity (Messick, 1989), the field of measurement, assessment, and testing has been calling out for a new and expanded evidential basis for test validation. Zumbo (2017, 2021, 2023a) responded to Messick's call by blending key ideas from construct validity theory and argument-based approaches that emphasize an explanation-focused view, transparency, and trending away from routine validation practices to shine a light on often hidden forms of test invalidity (see, also, Hubley & Zumbo, 2011, 2013; Zumbo & Chan, 2014a; Zumbo & Hubley, 2016). (p. 11)

[Figure 2](#) portrays a re-envisioning of a contemporary unified validity and validation framework, paying greater attention to the role of theory and values at each step, types of evidence included in construct validation, and the role of intended consequences and unintended side effects. It is an integrated conceptual framework for test validation that explicitly synthesizes construct theories and argument-based approaches.

To read and apply the framework depicted in [Figure 2](#), one would start at the far left of the figure with theories that define the attribute of interest and explicitly articulate its proposed uses (and ideally, what it should not be used for). One moves from left to right with a clear eye for when the loops double back. As Hubley and Zumbo (2013) state, their framework is consistent with Zumbo's (2009) view of validation as an integrative cognitive judgment involving a form of contextualized and pragmatic view of explanation – wherein explanation serves as a regulative ideal. Furthermore, their framework pays attention to the roles of values and theory at each step of validation, the types of evidence included in construct validation (see the large dashed circle at the center of the framework in [Figure 2](#)), and the role of intended consequences and unintended side-effects (concepts that they more fully introduce and explicate in their paper). Importantly, consequences and side effects of legitimate test use may also influence test score meaning, inferences, and decisions, which make them relevant to the validation process. Finally, in [Figure 2](#), the fact that some of the arrows loop back in the framework is particularly important, such that consequences and side effects of legitimate test use can affect the articulation of the construct. Likewise, we can see that the role of values is pervasive throughout the framework. [Figure 2](#) should not be seen as a radical departure from current validation theory and practices; it embodies, for the most part, contemporary thinking in the field (for more details, see Hubley & Zumbo, 2011, 2013).



**Figure 2.** Test validation framework depicting a synthesis of the explanation-focused view and argument-based approaches.



In Figure 1 earlier in this essay, in other sub-sections of this essay, and elsewhere (e.g., Hubley & Zumbo, 2011; Zumbo, 2017), we have argued that this sort of value-ladenness is already part of the science of measurement and testing -- and, we would argue, science more generally. Pretending that measurement and testing can be reformulated into value-free claims devalues perfectly good practices and stakes the authority of the science of measurement and testing on its separation from the community that enables and needs it. I am advocating an approach (one that we believe goes on regularly in basic and applied science) broadens our notion of objectivity and encompasses value-based decisions, such as those involved in test validation.

In an earlier section of this essay describing construct validity, it was noted that what has caused some confusion is that construct-valid tests provide information about (i) the study participant in terms of the construct and (ii) how the construct definition itself can be strengthened or extended and that questions of the theory of the phenomenon and its measurement cannot be answered independently of each other, and they co-evolve.

Distinguishing these two types of information and recognizing the importance of the second type features prominently in Hubley and Zumbo’s (2011, Figure 1) revised unified view of validity and validation. by a reciprocal feedback arrow from the "test score meaning, claims, and inferences" rectangle to the rectangle depicting the "psychological construct," and arrows in both directions between “test score meaning and inference” and intended (unintended) social and personal consequences. Hubley and Zumbo (2011) describe this as follows.

Our new model of validity and validation highlights several key features. First, one can envision that, based on a construct, one develops a test/measure to which one ascribes test score meaning and inference. From test-score meaning and inference emerge (a) intended social and personal consequences, but also (b) unintended social and personal side effects of legitimate test use. Unlike Messick, we argue there may be personal and social impacts. In addition, we think it is

helpful to use different terms to distinguish between intended consequences and unintended side effects. Importantly, consequences and side effects of legitimate test use may also influence test score meaning and inference, which. (pp. 225-226)

To make this less abstract, let us imagine that a researcher uses a self-report measure of academic self-efficacy to investigate if and how self-efficacy affects the meta-cognitive strategies of engineering students. A self-report measure of self-efficacy that is construct valid provides information about the respondents (engineering students) in terms of the construct; for example, students who possess a more profound understanding of self-efficacy are more successful in handling university-related tasks expected of engineering students and more effectively adopting learning strategies. Moreover, construct valid tests provide information about how the construct definition itself can be strengthened or extended, for example, whether the construct reflected in the self-report measure of self-efficacy is to be used, for example, with a different cultural group (e.g., Aboriginal peoples, international students) than the original test development target population whether a newly studied cultural group conceives of or values the construct, in the same way as the original group upon which the construct or measure was developed. This construct validity question asks how well a test or assessment travels through place and time, reflecting the degree to which the obtained scores reflect construct underrepresentation, construct-irrelevant variance, or both.

## 6. METHODOLOGICAL INNOVATIONS IN EXPLANATION-FOCUSED VALIDITY

It is important to distinguish between method and methodology at the outset of this essay section. Briefly, methods are means for helping us realize the objectives of our inquiry. In contrast, methodology contains resources such as the concepts and (formal or informal) logic for an informed understanding of our methods. An essential difference is that method is a component of methodology. However, methodology is more than just a collection of methods. The methodology provides the framework and the guidelines for conducting the research. As such, although there is some blurring of the distinction between method and methodology, this section tends toward the latter.

Haig (2019) provides a characterization and demarcation of method and methodology in the following.

It is important to distinguish at the outset between method and methodology. The term method derives from a combination of the Greek words *meta*, meaning following, and *hodos*, meaning the way, to give following the way, suggesting the idea of order. Applied to science, method suggests the efficient, systematic ordering of inquiry. The scientific method, then, describes a sequence of actions that constitute a strategy to achieve one or more research goals that have to do with the construction and use of knowledge. Researchers sometimes use the term methodology as a learned synonym for method (and technique). However, the term is properly understood as denoting the general study of methods and is the domain that forms the basis for a genuine understanding of those methods. To repeat, methods themselves are purportedly useful means for helping us realize chosen ends, whereas methodology contains the resources for an informed understanding of our methods. (pp. 528 – 529).

### 6.1. Third Generation DIF is About More than Just Screening for Problematic Items

#### 6.1.1. *Third-generation DIF led to methodological innovations*

Zumbo (2007b) outlined three generations of DIF research. The first generation explored the reasons for DIF in relation to test fairness and its concept formation. The second generation embodied the new terminology to develop statistical frameworks for DIF analysis. The third generation revisited the first generation and redefined DIF as arising from irrelevant factors of the item, the situation, or both, affecting the underlying ability and the test purpose. The inclusion of “situation” to the previous sources accounting for contextual variables to explain DIF, third generation DIF extended DIF theory and practice beyond the test structure, aligning

with an explanation-focused view of test validity that accounts for contextual sources of variation in item responses (Zumbo, 2007a, 2007b, 2009; Zumbo & Gelin, 2005). Thus, DIF can be meaningful and not just a nuisance for test interpretation and use.

Thus, the presence of DIF can be viewed as an opportunity to examine or explore the source of the differing probability of the groups endorsing the item. One may explore the source of DIF using cognitive interviews (Padilla & Benítez, 2014, 2017) or a latent class DIF model, which an ecological model can inform of item responding (Zumbo et al., 2015). This approach would potentially help inform the nature of DIF by drawing on information on underlying cognitive, psychosocial, or contextual processes during item response. In this sense, DIF becomes an assessment research, a validation method, and a window into response processes. This use of DIF methods is generally in agreement with the description of Cronbach and Meehl's (1955) and Loevinger's (1957) descriptions of construct validity as demonstrating that certain explanatory constructs account for performance on the test to some degree, Messick's (1989, 1995) substantive validity, and more directly to the ecological model of item responding described in, for example, Zumbo et al. (2015).

Zumbo and Gelin's conceptual framework is the precursor to the ecological model of item responding (Zumbo et al., 2015), which in educational assessments can include items and test characteristics, individual, classroom, or school characteristics, and country factors. Importantly, as described by Zumbo (2007b), Zumbo's (2007a) explanation-focused view of validity, DIF becomes intimately tied to test validation, not only in the sense of test fairness. Zumbo (2007b) describes one purpose of third-generation DIF: trying to understand item response processes. In this use, DIF becomes a method to help understand the cognitive and psychosocial processes, or both, of item responding and test performance and investigating whether these processes are the same for different groups of individuals. In this use, DIF becomes a framework for considering the bounds and limitations of the measurement inferences.

The central feature of this view is that validity depends on the interpretations and uses of the test results and should be focused on establishing the inferential limits (or bounds) of the assessment, test, or measure (Zumbo & Rupp, 2004). In short, invalidity distorts the meaning of test results for some groups of examinees in some contexts for some purposes. Interestingly, this aspect of validity is a modest but significant twist on the ideas of test and item bias of the first-generation DIF. That is, as Zumbo (2007a) and Zumbo and Rupp (2004) noted, test and item bias aim analyses at establishing the inferential limits of the test—that is, establishing for whom (and for whom not) the test or item score inferences are valid.

### ***6.1.2. The ecological model of item responding and subtest or test performance as a methodological innovation***

As Zumbo (2018b) noted, over a 20-year period of normal development starting in 1985, the initial enthusiasm for DIF research began to wane in the field. However, beginning in 2005, developments in third-generation DIF, mixed-methods DIF, explanation-focused validation studies, DIF informed by an ecological model of item responding, latent class DIF, and the use of DIF in response processes validation research have led to a resurgence of interest in DIF and this emerging paradigm shift.

This renewed enthusiasm for DIF research led to new psychometric statistical methods by myself and others founded on a recognition that (i) the investigation of DIF is important for any group comparison, diagnosis, or classification based on assessments or surveys because the validity of the inferences made from scale scores could be compromised if DIF is present (e.g., Li & Zumbo, 2009; Rome & Zhang, 2018), and (ii) identifying the determinants (or explanatory theory) of item and score variation is central to a strong theory construct validity (Messick, 1995; Zumbo, 2007a, 2009). Regarding explanatory DIF, knowing why, how, and what

---

mediates, moderates, or functions as a mediated-moderator (Wu & Zumbo, 2008) of item responses bridges the inferential gap from test scores to claims about constructs and provides an understanding and description of the enabling conditions for item responses (Zumbo et al., 2015).

A large part of this richer explanation provided with the emerging paradigm shift stems from what I refer to as embracing the many ways of being human in assessment research and test validation, which, in the current discussion, implies as described by Zumbo et al. (2015), that there is a tendency to treat grouping variables for DIF analyses as what philosophers would describe as *natural kinds* (Kaldis, 2013). In our context of DIF analyses or validation studies of group differences, a type of natural-kind essentialism is often unknowingly invoked wherein grouping variables are interpreted as reflecting intrinsic or essential features that correspond to the real, mind-independent groupings in nature and are characterized by shared essences. This approach is motivated by practices in the natural sciences; however, there is little evidence for doing so in the educational and psychological measurement field. Several recent DIF studies give passing recognition that there may be an inherent heterogeneity in these grouping characteristics and that these grouping variables or categorizations reflect historical categorizations or some human interests or purposes, which are referred to as social or human kinds by Kaldis and others. However, for the most part, assessment research, DIF studies, and validation practices continue to mirror the natural sciences, unknowingly invoking natural kinds incorrectly. Situating the many ways of being human at the center of my explanatory-focused view urges these researchers to question these practices.

### ***6.1.3. An attempt to clarify the terminology: Is it situation, ecology, and context, or a subset of them?***

Unlike Zumbo (2007b), in Zumbo et al. (2015), testing situations are deemphasized in favor of the richer concept of human ecology, and we speak of contexts periodically. As emphasized in biopsychosocial theories (e.g., Bronfenbrenner, 1979, 1994), ecological conditions shape and promote psychological development and growth. These conditions include home, school, and workplace environments. Building on such ecological theories, Zumbo and colleagues (2015) described the ecology of item responding with the item responding embedded in a multiplicity of contexts. Views of measurement validity by Messick, Zumbo, and others focus on evidence about why and how people respond as central evidence for measurement validation. In line with Messick's (1989, 1995) articulation of substantive validity, the ecological model of item responding provides a contextualized and embedded view of response processes conceptualized as a situated cognitive framework for test validation (Zumbo, 2009; Zumbo et al., 2015).

Earlier research by my colleagues and I did not satisfactorily distinguish between situation, ecology, and context, often choosing to use them interchangeably. For example, Zumbo and Gelin (2005) state that the ecology of item responding allows the researcher to focus on sociological, structural, community, and contextual variables and psychological and cognitive factors as explanatory sources of item responding. We hope this broad treatment would be most beneficial to further the use and development of our novel ecological model of item responding in assessment research and test validation. After all, a large body of psychological research dating back to the 1970s continues struggling to differentiate these terms adequately. Where there is common practice, it is local to a particular research topic. For example, the *Journal of Personality* devoted a special issue to personality and its situational manifestations, bringing together personality, social, self, clinical, and cultural psychologists who have attempted to contextualize the self, personality, attachment, and cultural constructs in an integrative fashion (Roberts, 2007).

One can take the lead from Yang et al. (2009) if one wishes to distinguish between situation, environment, and context, along with Zumbo et al.'s (2023) description of the assessment



encounter in the typical testing or assessment setting as an item playing the role of a stimulus in a traditional behavioral stimulus-response (S-R) language. Yang et al. state that situation and related concepts, such as stimulus and environment, are used interchangeably to refer to the external conditions surrounding human activities. They provide a distinction as follows.

“... [the] situation differs from the other two in both the levels of analysis and disciplinary foci. In terms of levels of analysis, situation is typically conceptualized at the intermediate level, while stimulus is at the micro level concerned with a specific object that gives rise to the organism’s response (Sells, 1963), and environment is at the macro level concerned with the aggregate of larger physical and psychological conditions that influence human behaviors (Wapner & Demick, 2002). Thus, the concept of situations can be considered at the level between stimulus and environment, such that a stimulus may be a part of a situation, and a situation may be a part of the environment.” (p. 1019)

In terms of context, Bazire and Brézillon (2005) describe it as superordinate to the environment and has “... two dimensions: (1) Ecology: aspects of the school that are not living, but nevertheless affect its inhabitants (resources available, policies and rules, and size of the school); and (2) Culture to capture the informal side of schools.” (p. 38)

Therefore, as a tentative way forward, if one wishes to distinguish these levels, we have a stimulus, an item or task on test or assessment, situation, environment, and context or environment and context undifferentiated. In this description, context includes dimensions of ecology and culture.

## **6.2. An Entrée for Embracing the Many Ways of Being Human in an Explanation-Focused Framework**

Zumbo (2007b) defined the third-generation DIF as investigating why DIF occurs. Unlike the first and second generations, in the third, Zumbo and colleagues (Zumbo et al., 2015; Zumbo & Gelin, 2005) expanded beyond item characteristics, such as differentially unfamiliar terminology, to understand the item responses. Their expanded explanatory sources included psychological and cognitive factors, physical and structural settings of the community, and the social context that needs to be explored.

It is important to remember that we adhere to the view that neither the test taker nor the cognitive processes in item responding are isolated in a vacuum. Instead, test takers bring their social and cultural present and history to test taking. We accept as our starting point the widely received view in the broader social sciences that human beings have evolved to acquire culture from birth and that the culture to which an individual is exposed, and the ecology of their lives, affects their basic psychology and cognition, including, in our case, item responding. As such, this ecological view of item response or test performance rests on an evolutionary, adaptive view of human beings in continuous interaction with their environment, particularly considering measurement validity and response processes.

When viewed within this ecological framework, item responses and test performance cannot be simply attributed to the individuals or the environment but to the relationship between the two. In so doing, one can move to a contextualized form of explanation that embraces the many ways of being human and works against a binary structure of variables considered of a natural kind that explain test performance. That is, in describing their novel ecological model of item responding, Zumbo et al. (2015) further motivate the important role of the many ways of being human as follows:

In short, Third Generation DIF is part of building an ecological model of item responding and assessment. The ecology of item responding, as Zumbo and Gelin (2005) note, allows the researcher to focus on sociological, structural, community, and contextual variables, as well as psychological and cognitive factors, as explanatory sources of item responding and hence of DIF (Zumbo & Gelin, 2005). (p. 139)



Nevertheless, there is tension between the aspirations of an equitable and socially just assessment and validation methodology where they are used in education and psychology settings and the realities associated with its implementation. Zumbo et al. (2015) characterize this tension as follows:

For example, a classical example of DIF studies includes a focus on gender-related DIF. However, gender has, in the main, been characterized in the binary as biological sex wherein (binary) biological sex differences on item performance that are eventually explained by item characteristics such as item format and item content. In Third Generation DIF “gender” more properly should be considered a social construction, and gender differences on item performance are explained by contextual or situational variables (ecological variables, if you wish), such as institutionalized gender roles, classroom size, socioeconomic status, teaching practices, and parental styles. We believe that these richer ecological variables have been largely ignored in relation to explanations for (and causes of) DIF because of the focus on test format, content, cognitive processes, and test dimensionality that is pervasive in the second generation of DIF. (p. 139)

As such, the many ways of being human embodied in the third-generation DIF “gender” more properly should be considered a social construction, and gender differences in item performance are explained by contextual or situational variables (ecological variables, if you wish), such as institutionalized gender roles, classroom size, socioeconomic status, teaching practices, and parental styles. What is noteworthy is the shift from considering gender differences as a nuisance variable in the interpretation of the item and test score to explanation-focused attention, where gender as personal identity plays a role in helping us understand the process of item responding. In this example, the subtle turn to focusing on the encounter of the test taker and the assessment or item includes what anthropologists would describe as a performative component and social encounter (Maddox et al., 2015; Maddox & Zumbo, 2017; Zumbo, 2007a).

Aligned with the turn to focusing on the encounter of the test taker and the assessment or item, I prefer the lens of the many ways of being human rather than the more conventional concept of fairness for three reasons. First, the former fosters a more expansive view than fairness, *per se*, because it urges the assessment researcher to abandon the notion of demographic variables as reflecting natural kinds. Second, it positions the assessment researcher to consider test taking as an encounter similar to the description above rather than a static contrived space depicted in my contrasting *in vivo* compared to *in vitro* assessment settings. Third, while there is a general belief in educational and psychological measurement that fairness is a fundamental validity issue that should be addressed right from the beginning of the test development process, the term fairness has no single technical meaning. It is used in many different ways in the field. As I highlighted in my description of the Draper-Lindley-de Finetti (DLD) inferential framework (Zumbo, 2007a), if we are to interpret test scores fairly, they must be comparable for all individuals in the population that the test aims to measure. It is also important that the scores are not influenced by factors that are not relevant to the construct we want to measure.

Linking DIF to the broader issue of measurement validity, the ecological model further articulates what “context” means in Zumbo’s (2009) view of validity as a contextualized and pragmatic explanation—that is, the multilayered ecology is the context. Furthermore, by accounting for contextual variables to explain DIF, third-generation DIF is aligned with an explanation-focused view of test validity that accounts for contextual sources of variation in item responses (Zumbo, 2007b, 2009). Finally, the ecological model is a foundation for the statistical and psychometric methodology of item responding. Explicit consideration of social and personal consequences and side effects might enlighten us concerning whether personal (e.g., age, gender, culture) and contextual factors (e.g., learning environment, social support, gender socialization) are part of the construct of interest or external to it (Zumbo, 2015).

### **6.3. The Importance of, and Multiple Ways to Think About, Loevinger's Two Test Validation Settings**

Zumbo (2017) describes an ecologically informed in vivo view of validation practices centering on response processes assessment research in a paper of the same title as this section. Trending away from routine procedures toward an ecologically informed in vivo view of assessment research and validation practices invokes what Zumbo (2015) refers to as an in vivo view of testing and assessment rather than the more widely received in vitro view. Doing so, I would argue, necessitates an ecological model of item responding and test performance (Zumbo et al., 2015). The ecological (situated) point of view is tied closely with the notion of in vivo. As Zumbo (2017) states, therefore, when adopting Zumbo's explanation-focused, ecological, and in vivo approaches, there is a rhetorical move from how the environment affects the person to a type of interactivism in which the test taker is situated within these enabling conditions and highlights processes and forms of influence of the context/situation (sometimes referred to as the environment) on the test taker that is obscure or entirely absent from the received standard view of item and test responding.

Those who investigate the validity of inferences drawn from assessment practices are said to engage in validation research. "The process of validation involves presenting evidence and a compelling argument to support the intended inference and to show that alternative or competing inferences are not more viable" (Hublely & Zumbo, 2011, p. 219). The in vitro versus in vivo contrast clarifies a remark by Loevinger (1957), referred to earlier in this essay as one of Loevinger's ideas that has been generally overlooked in the test validity research literature. Loevinger recommended that two basic contexts for defining validity be recognized, administrative and scientific, which in my language would be in vitro and in vivo, respectively. According to Loevinger, there are essentially two kinds of administrative validity: content and predictive-concurrent, whereas there is only one kind of validity that exhibits the property of transposability or invariance under changes in an administrative setting, which is the touchstone of scientific usefulness: construct validity (Loevinger, 1957, p. 641). Another way of describing this is gathering test validity evidence when an assessment is designed and developed in a controlled setting that we can describe as in vitro for use in the intended context(s) and populations. Loevinger's scientific context of test validity and assessment evidence drawn from the diverse and varying contexts of assessment use reflected the many ways of being human.

Related to these two settings for validity studies, during my description of in vivo and in vitro views of validation practices, I also introduced the "off-label" use of a test or assessment (Zumbo, 2015). I described off-label use as including test administration in an unapproved or undocumented manner during the administrative validation setting. Off-label use may also include using a test or assessment for an unapproved purpose, for an undocumented or unapproved intended target test group such as an age group or a cultural group. Generally, I would caution against off-label use, but there is some subtlety. There is not a great deal of discussion of off-label use of tests, partly because many (but not all) test developers do not necessarily want to dictate the use of a test, or more specifically, what it should not be used for, or to whom the test should be administered. Some test developers are better at this than others. However, there are many cases where "off-label" would be difficult to determine because "on-label" is not clearly articulated and documented.

It is nearly impossible to police off-label use, and it likely happens often. Test users may use a test off-label; however, off-label use must better serve the student (or, more generally, the test taker) than other test alternatives, such as no test information or a test already known to be inappropriate. In addition, the off-label use must be supported by evidence or experience to support the lack of unintended negative consequences and efficacy in construct interpretation. I will close my remarks with a word of caution that off-label use may alter the construct or lead

to (a) intended social and personal consequences and (b) unintended social and personal side effects of off-label test use (Hubley & Zumbo, 2011).

Regarding how observations of real-life testing situations can provide insights into test validation, O'Leary et al. (2017) raise several points about the differences between intended and actual interpretation and use of scores. These points are of the utmost importance when considering Loevinger's two basic contexts for defining validity, administrative and scientific, and my description of in vitro and in vivo assessment research and validation settings. The essence is that test validation research only conducted in idealized (administrative or in vitro) settings may not address the central question of construct validity in the wild, in vivo. Framing their argument from Hubley and Zumbo's (2011) framework for considering consequences for test interpretation and use, O'Leary et al. make the case that "... when there is an alignment between intended and actual interpretations and use, then the purpose of tests, the intended personal and social consequences at the core of assessment practice, have the greatest chance of being realized" (p. 16). Furthermore, O'Leary et al. remind the reader that validity is about both interpretations and use of scores; however, in addition to the known and anticipated interpretations and use of scores, many unknown interpretations and use are comprised of off-label, unintended and/or potentially illegitimate use and users of test scores (Zumbo, 2015). Although certain views of validity set aside concerns about test interpretations, use, and consequences (Borsboom et al., 2004, 2009), O'Leary et al. make the point most convincingly in the following.

Essentially, at its very core, validity is about the interpretations and use that are based on test scores as opposed to the actual testing instrument itself (Hubley & Zumbo, 2011) and, of equal importance, it must be evaluated with respect to "the purpose of the test and how the test is used" (Sireci, 2009, p. 20). (p. 17).

Currently, validation is concerned with providing theory and evidence in support of intended or proposed interpretations and use. However, the importance of providing evidence for how users make inferences and take actions has recently been recognized (Hattie & Leeson, 2013). Nevertheless, within the Standards, there is no clearly articulated form of validity evidence or guidelines related to a consideration of linking how test score users make actual interpretations and subsequently plan uses based on scores. This presents a challenge. (p. 18)

[D]espite much movement in validity theory, validity in practice is dominated by whether a test is capable of achieving its stated aims. This is disappointing. If validity is to be truly concerned with the appropriateness of interpretations and use, then evidence of the quality, appropriateness, and effectiveness of the actual interpretations that test score users make and the actions they plan based on how scores are reported must be central to both the validity and validation processes. Not only would this result in a more authentic realization of the current definition, but consideration of such evidence could help to improve the overall quality of the outcomes of testing by (1) helping to identify poor interpretations and uses, unanticipated interpretations and uses, and misuse before the fact, and (2) subsequently informing necessary improvement with regard to how scores are being reported. (p. 19)

#### **6.4. Response Processes Are Important to Test Validation: Insights from a Broadened View**

Let us remind ourselves that whether one considers psychometric test theory or design more generally, a basic building block of any test or assessment is the encounter or what the mathematically oriented test theorist would describe as the interaction of a test-taker and an item or task. This encounter results in a response scored as correct/incorrect or for partial points and a composite score across the items computed for knowledge or achievement tests. On the other hand, a psychological test or measure may be viewed as a set of self-report questions (also called "items") whose responses are then scored and aggregated in some way to obtain a composite score. In many psychological measures (e.g., attitudinal measures), there are no

“correct” or “incorrect” responses, per se. Therefore, what is scored are compelled self-report responses.

It is important to note that I foreground the encounter of a test-taker and item or task, sometimes called the interaction of a test-taker or respondent with an item or task. This encounter or when a test taker interacts with an item or task is paramount in my explanation-focused view. However, the product or outcome of this encounter is the focus of what is to be explained in explanation-focused validity. This point of paramount significance is captured in the language of measure-theoretic mental test theory, as described earlier in this essay, as the “measurement process.”

Zumbo et al. (2023) recently described a broadened view of response processes focusing on informing validation practices. As highlighted therein, although response processes are often listed as a source of validity evidence, we rarely see a clear conceptual or operational definition of response processes; rather, the focus is on the techniques and methods. As such, method trumps clear definitions, and, as a field, we continue to conflate method and methodology—much like we conflate validity and validation. This focus on technique and methods is not to say that, as described in detail by Zumbo et al., important definitions have not been offered in the field.

Zumbo et al. present a broad definition that expands the evidential basis to include methods such as response times, eye tracking methods, mouse clicks, keeping records that track the development of a response, analyzing the relationship among components of a test or task or between test scores and other variables that address inferences about what they describe as product and process constructs. Zumbo et al. (2023, Figure 1) depict the space between a test question or task presented to the test-taker and when they respond, highlighting response processes and process data, highlighting the context of computer-based testing. However, the description holds for paper and pencil exam delivery. In behaviorist language that shaped early assessment and testing theories, this test question or task is described as the “stimulus (S).” The response to the item or task is the response (R) in that stimulus-response (S-R) view of behavior and response processes happen in the space between S and R. Cronbach and Meehl (1955) acknowledge this S-R space by invoking earlier concepts of intervening variables and hypothetical constructs (MacCorquodale & Meehl, 1948). The later information processing and cognitive psychologists conceived this space as holding the mental processes. To access this space, Messick referred to mental probes (e.g., think-aloud methods).

The Zumbo and Hubley (2017) volume offers a broadened view of response processes as mechanisms explaining what people do, think, or feel when interacting with and responding to items. Thus, response processes go beyond cognition, including emotions, motivations, and behaviors affecting item and test score variation. Zumbo et al. (2015) propose an ecological model of item responding that considers contextual influences from the test takers’ lived experience, family setting, and larger community or national characteristics (Chen and Zumbo, 2017; Woitschach et al., 2019). Finally, building upon developments by Maddox, Zumbo et al. characterize this space as temporal, cognitive, affective, physiological, embodied, and material features.

The essential differences between the theories and viewpoints described above reflect the breadth and scope of characterizations of response processes and the terrain of future research. Some early views conflated what response processes are with how they are attained. For example, Messick characterizes response processes arising from mental probes. Other theories conceive of response processes as mostly cognitive and physiological, wherein the intervening variables are the unobserved mechanics of the *process* leading to the response.

Zumbo et al. conclude as follows.

Our proposed holistic framework, therefore, articulates a definition and relation between test constructs and process constructs, highlighting the need to rigorously conceptualize and validate the way that response processes and “process data” are treated as measurement opportunities. (p. 259)

### 6.5. Test Validation as Jazz

It is important to remind ourselves of three points about test validation. First, no widely accepted series of steps can be followed to establish the validity of the inferences one makes from measures in the varied and disparate fields wherein measurement is used. Having said this, however, it is important to note the distinction I make between validity, per se, and the process of validation. I consider validity to be the establishment of an explanation for responses on tasks or items – the emphasis being inference to the best explanation as the governing aspect. The validation process informs that explanatory judgment, hence, by nature, brings the validation process squarely into the domain of disciplined inquiry and science.

There are many metaphors discussed in the literature for the process of validation: (a) the stamp collection, (b) chains of inference, (c) validation as evaluation, and (d) progressive matrices, to name just a few. Zumbo (2007a) described his vision of assessment research and test validation as jazz – as in the musical style. With validation as jazz, I principally borrowed the tenets of sound coming together, but that the coming together is not necessarily scripted. All sorts of notes, chords, melodies, and styles come together creatively (including improvisation that is particular to that one song or performance) to make music. Perhaps the same applies to the process of validation: no one methodology or script can be applied in all assessment contexts.

Maddox and Zumbo (2017) riffed on Zumbo’s (2007a) idea that test validation is like jazz. They set the tone for their description of response processes as evidence for test validity as follows:

Think aloud protocols are considered by some to be the received method for investigating response processes from an individual cognitive perspective. In contrast, we consider real-life testing situations as distinctive social occasions that merit observation (Maddox, 2015). While testing situations reveal observable structures and patterns of behaviour, every performance is somewhat different. Like jazz, investigating the testing situation involves elements of improvisation. We see our task as to listen to those patterns and improvisations. That is, to hear music rather than noise. (p. 179)

By focusing on observations of interaction in face-to-face testing situations and the character of improvisations, Maddox and Zumbo expand the set of information available to understand and explain response processes (see Zumbo, 2007a, 2007b; Zumbo et al., 2015). Maddox and Zumbo go on to unpack this further in the following.

However, our aim is not simply to amplify individual differences in test behaviour. Instead, by observing the testing situation we hope to identify clues about the way the test is constructed, understood, and performed as a social occasion. This may include, for example, observation of interaction within wider social structures or social relations that inform and mediate assessment performance. These act as enabling conditions for the abductive explanation for variation in test performance. (p. 180)

In terms of the process of validation (as opposed to validity, itself), the methods described herein work to establish and support the inference to the best explanation—i.e., validity itself; so that validity is the contextualized explanation, whereas the process of validation involves the myriad methods of psychometrics, including what we call “psychometric-ethnography” (Maddox et al, 2015). Zumbo’s abductive approach to validation seeks the enabling conditions through which a claim about a person’s ability from test performance makes sense (Stone & Zumbo, 2016; Zumbo, 2007b, 2009). (p. 180)

They employ the rhetorical device of testing in vivo, described earlier in this essay, to capture



the process of interaction and social embeddedness of the testing situation that mediate and shape individual test-taker response processes). As they state, although it may not be considered construct-relevant by some assessment researchers, such ecological information provides a potential explanation for variation in response processes rather than being considered a source of pollution or cultural noise to be controlled and excluded. The contrasting idea is that assessment practice and explanation could somehow occur “in vitro,” as if isolated from its cultural and ecological setting and sources of influence that occur in real-life operational contexts.

Maddox and Zumbo take the assessment research and test validation as jazz metaphor one step further by focusing on the dynamics of interaction in testing situations (e.g., see Maddox, 2015) while recognizing the potential for those interactions and responses to be influenced by larger-scale “off-stage” (Goffman, 1959, 1964) dimensions of the testing situation such as social institutions, social relations, norms, and beliefs that we might associate with Zumbo et al.’s ecological model.

### **6.6. Test-Taker-Centered Assessment and Testing and Test Validation as Social Practice: The Case of Inclusive Educational Assessment, Neurodiversity and Disability**

Validation research in support of claims made from assessments in the twenty-first century has become more nuanced and less formulaic due in considerable measure to the field of assessment embracing, rather than merely accommodating, the diversity of test takers. Several assessment theorists have taken on the challenges (and the promise) provided by awareness and, hopefully, greater understanding and respect for test takers who represent neurodiversity, diverse cultures, beliefs, and historical experiences. Within this context, my explanation-focused ecologically shaped in vivo view of validation practices embracing the many ways of being human has developed over the past decade. Although precursors to this approach date back to the early 1970s, the expansion of this research model became possible more recently with digital innovations and advances in data science. This section of the essay will briefly describe the motivation for and critical concepts in this assessment design, validation, and research model to address the call for greater attention to inclusive educational assessment, neurodiversity, and disability.

Zumbo et al. (2023) highlighted that disabilities and neurodiversity can lead to test takers responding to test items in ways that deviate from established models. They state that human neurobiology has a broad diversity; the human brain develops and functions in countless ways, resulting in a test-taking population with diverse strategies and responses; therefore, there is a need to recognize that, rather than anomalies, test-takers with disabilities and learning differences represent a sizeable minority (p. 257).

A central tenet of the test validation and assessment research method, which I introduce herein, is that engaging with test-takers of a range of neurodiversity and disabilities to learn about their experience and insights into test design, administration, and interpretation of test scores is a tremendous step forward. However, as Addey et al. (2020) highlight, it is uncommon to situate psychometric measurement validation research within a context where respondents, caregivers, or families engage as partners in the psychometric validation process. Mobilizing knowledge from strategies in health and human development, I propose that educational assessment take up a test-taker-centered assessment and testing framework that theoretically centers on Addey et al.’s test validation as social practice.

Test-taker-centered educational assessment and testing are driven by test takers and members of their extended support systems’ expressed values, preferences, and needs. It involves partnering meaningfully with test takers and members of their extended support systems to decide what educational constructs to assess, how to assess them, how to integrate these various

constructs into a profile (rather than an aggregate construction), who should get the results, and how to use those results. I recognize this is not feasible in all educational settings; ideally, assessment design, delivery/administration, scoring, interpretation, and reporting of the outcomes are test-taker-driven and co-created. Furthermore, ideally, the educational assessment data are the property of the test takers and members of their extended support systems.

Zumbo (2023c) recently addressed the urgent call that brought testing and assessment specialists, educators, and policy researchers to the 2023 “Cambridge Symposium on Inclusive Educational Assessment, Neurodiversity, and Disability.” My central message is that as a discipline, we must reorient our validation practices and open the test design, delivery, and validation process to diverse voices and contributions beyond our typical disciplinary focus. Addey et al.’s (2020) framework of test validation as social practice can help bring attention to the principal challenges and opportunities of inclusive educational assessment, neurodiversity, and disability.

The challenges and opportunities of inclusive educational assessment, neurodiversity, and disability are an ideal space to implement Addey et al.’s description of co-construction and democratic engagement of diverse members of the test-taker and stakeholder populations. In short, Addey et al. (2020) consider the socio-material validation practices of assessment actors as they assemble validity with the explicit goal of “creating a democratic space in which legitimately diverse arguments and intentions can be recognized, considered, assembled and displayed” (p. 588). As a social practice, “assembled validity” suggests that validity arguments are assembled iteratively in dialogue, as validation evidence is identified and collected, and new actors are enrolled.

The task of democratically assembling validity would be to identify and reconcile (rather than ‘rebuff’) the plural and legitimate theories of different stakeholders (their epistemologies and contexts). Central to this democratic engagement are principles of (true) consultation and duty to consult modeled upon the Duty to Consult with First Nations Peoples Sec. 35 Canadian Constitution and the United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP). This meaningful consultation should have the following features.

- Test developers and, where appropriate, policy specialists have a duty to consult experiential experts, that is, test-takers (or their guardians) who reflect the range of neurodiversity and disability in the target population when contemplating conduct that may have an adverse effect on them.
- An essential feature of this consultation is information sharing and an eye to resolving potential adverse impacts identified by the ‘experiential experts’ (Zumbo, 2016).
- It entails listening to and accommodating concerns, being willing to amend test design proposals in the light of information received, and providing feedback.
- A dialogue must ensure that it leads to a demonstratively serious consideration of experiential experts’ requests – no “faux consultation.”

Importantly, the scientific interest and the duty to consult do not operate in conflict. This form of (true) consultation describes a fundamentally different relationship with the community of test-takers, leading to critical test-taker-oriented testing and assessment practices.

## 7. CONCLUSIONS

To set the tone of this closing section, a statement from the first section of this essay bears repeating. As Zumbo and Chan (2014a) show via a large-scale meta-synthesis of the genre of reporting test validity studies across many disciplines in the social, behavioral, and allied health sciences, this research is largely uncritical in presenting their subject matter, rarely indicating what of many possible validation frameworks were chosen nor why (Shear & Zumbo, 2014). As hidden invalidities may undermine test score claims, this research should focus on the

concept, method, and validation process since invalid measures may harm test takers.

As we observed in the introductory section of this essay, the late 20th and early 21st century saw a global increase in the use of assessments, tests, and instruments in the social sciences based on educational and psychological measurement developments that coincided with a growing economy of global assessment and testing. Rapid assessment theory and practice changes during this period left some important issues unresolved or in the background.

The essay is divided into two parts. The first part, comprised of sections two and three, described the organizing principles that allow me to catalog and then contrast the various implicit or explicit definitions of validity and then report on a novel historical analysis addressing whether and, if so, what progress has been made in validity theory since the early 1900s. A meta-level theme emerged, reflecting a trend in explanation-focused theories of test validity. Along the way, I highlighted the context of the intellectual and commercial forces that shaped the changes in test design, development, and delivery and the changes in validity theory since the mid-1950s but focusing on developments since the mid-1970s, pointing to possible hidden invalidities. Building on the outcome of the first part of this essay, the second part, comprised of sections four through six of this essay, presented the primitives and settings that fostered the development, a detailed description of, and the innovations on the horizon in validation methods related to my explanation-focused view of test validity and validation methods.

These two sections of the essay draw to the foreground what Zumbo (2019) describes as the tensions, intersectionality, and what is on the horizon for assessments in education and psychology. As we saw in sections two and three of this essay, by the 2020s, the dominant theoretical views of validity aimed to expand the conceptual framework and power of the traditional view of validity established in the first fifty years of that century. Of course, it is important to note that there is nothing inherently wrong with the conventional views of validity that appeared in the first 60 years of the 20th century; however, hidden invalidities that are not considered in the first four definitions of the concept of validation may undermine test score claims.

Developers and purveyors of tests and assessments, those employed and profiting from the testing and assessment industrial complex, desire to ensure that their assessment tools and delivery systems are grounded in our most successful psychometric and statistical theories. They aim to do social good while serving their economic and financial imperatives. This goal is not necessarily untoward or ignoble; Zumbo (2019) describes a social and economic phenomenon reflecting financial globalization and international competitiveness. There is a notable increasing desire of those of us outside of the test and assessment industrial complex, per se, to ensure that the philosophical, economic, sociological, and international comparative commitments in assessment research are grounded in a critical analysis that flushes out potential invalidities and intended and unintended personal and social consequences. It is evident from the changes in validity theory and validation practices that these two strands are not necessarily working in opposition but are connected by a common body and goal of increasing the quality of life of our citizens globally.

Let me now turn to several observations and key messages from this essay. First, it is important to note that following the historical analysis in sections two and three of this essay, I identify the locus of the theoretical commitment of my test validity's commitments not in appeals to scientific theory in the sense used by several other validity theorists through the history of the topic, but in explanation of variation in item and test performance. As demonstrated throughout this essay, I ground my appeals to explanation in philosophical theories of scientific explanation. One reason to appreciate this richer, philosophically-informed cognitive view of explanation is that it has implications for my heterogeneity hypothesis— perhaps, more

---

accurately described as a hypothesis that does not prioritize homogeneity of the response process and validity evidence.

Second, I cannot stress this enough: from my point of view, assessment research and validation that embraces the many ways of being human aim to identify and explain sources of variation in test response processes that are endogenous to the testing situation and that lie outside “individual” notions of cognition (Zumbo et al., 2015). An explanation-focused view of validity with an ecological model of item responding allows a researcher to focus on anthropological, political, sociological, structural, and community and contextual variables and psychological and cognitive factors as explanatory sources of item responding (Zumbo et al., 2015). The ecological (situated) point of view is tied closely with the notion of *in vivo*. Therefore, when adopting Zumbo’s explanation-focused, ecological, and *in vivo* approaches, there is a rhetorical move from how the environment affects the person to a type of psychosocial interactivism in which the test taker is situated within these enabling conditions and highlights processes and forms of influence of the context/situation (sometimes referred to as the environment) on the test taker that is obscure or entirely absent from the received standard view of item and test responding.

Third, Zumbo (2005, 2007a, 2009) described an explanation-focused approach to test validity in which test validation centrally involves making inferences of an explanatory nature, highlighting inference to the best explanation (IBE). This reliance on explanation and IBE was presented contra the dominant mode of construct validation framed as hypothetico-deductive empirical tests in line with Cronbach and Meehl and those scholars who advocated that view. My view of test validity is also meant to guide our assessment research and reflects my perspective that validity: “[e]xplanation acts as a regulative ideal; validity is the explanation for the test score variation, and validation is the process of developing and testing the explanation” (2009, p. 69).

Fourth, as described in Zumbo et al. (2015) and Zumbo (2017), it bears repeating that my explanation-focused view of validation and assessment research adheres to the view that neither the test taker nor the cognitive processes in item responding are isolated in a vacuum. Instead, test takers bring their social and cultural present and history to test taking. We accept as our starting point the widely received view in the broader social sciences that human beings have evolved to acquire culture from birth and that the culture to which an individual is exposed, and the ecology of their lives, affects their basic psychology and cognition, including, in our case, item responding. As such, this ecological view of item response or test performance rests on an evolutionary, adaptive view of human beings in continuous interaction with their environment, particularly considering measurement validity and response processes.

Fifth, when viewed within this ecological framework, item responses and test performance cannot be simply attributed to the individuals or the environment but to the relationship between the two. In so doing, one can move to a contextualized form of explanation that embraces the many ways of being human and works against a binary structure of variables considered of a natural kind that explain test performance. That is, in describing their novel ecological model of item responding, Zumbo et al. (2015) further motivate the important role of the many ways of being human.

Sixth, drawing a thread from what led up to the first description of the explanation-focused view in my Messick Award Lecture (Zumbo, 2005) to my earliest descriptions (Zumbo, 2007a, 2007b, 2009; Zumbo & Gelin, 2005) allows for a fuller description of what I see on the horizon of assessment research and test validity from the vantage point of my explanation-focused view and discuss the ideas that influenced it and its statistical methods and share my reflections, critiques, and queries on its development. Likewise, in the last three sections of this essay, I describe how the current version of my explanation-focused view of assessment research and

test validation responds to the global rise of assessments since the late 20th century coincided with a period of rapid development and increased availability of computational sophistication.

Seventh, basing validation research on a coherent theory of validity and aligned validation methods that incorporate the many ways of being human is the central issue in addressing the tensions described at the start of this essay. Developments in educational and psychological measurement theory and methodological innovations: The trend to more elaborated views of validity and validation. Since the mid-1950s, the dominant modes of discourse: (a) Cronbach and Meehl (1955) was the key point where until Messick took the mantle, major developments were in response to Cronbach and Meehl and, most recently, Kane and (b) post-Cronbach and Meehl, the trend in theorizing has been in terms of what Zumbo (2009) describes as explanation-focused approaches.

Eighth, as assessment researchers, we want to know why different test takers often respond differently to the same test question or task. The aggregate score of the item responses results in a test score that displays variation across individuals. Suppose one asks oneself why this research question seems so pressing. In that case, I think the answer must be because, much more often, we use tests and assessments under the assumption that there are interpretable differences across individual test-takers regarding the psychological attribute we intended to measure with the test. Consequently, non-uniform, unexpected, unplanned phenomena confront us as anomalies. That is, perceived anomalies are necessary conditions of scientific research. When nothing is regarded as strange and unaccounted for, nothing is regarded as in need of explanation. The perceived necessity for an explanation of something is the threshold of scientific investigation.

The ninth and final remark is that to address the tensions I described in the opening section of this essay and the expanding diversity of test-takers and testing settings, the next generation of assessment researchers must possess the following.

- The next generation of assessment researchers needs to be fluent in validity theories and aligned validation practices and appreciate how the discipline's history both binds us to a narrow tradition and potentially liberates us to face unanticipated challenges from within and from outside of the discipline, including social changes.
- The next generation needs to recognize that initially, classical test theory seems simple. However, its description and interpretation have changed over time. Interpreted as conditioning on all possible outcomes of the measurement process  $X$  for a particular test-taker, the variation in observed test-taker scores includes measurement error and variation attributable to the different test ecological testing settings. As such, it is now aligned with the explanation-focused view wherein item and test performance are the object of explanatory analyses.
- Therefore, the next generation needs to appreciate the new re-interpretation of a true score afforded by measure-theoretic mental test theory; true scores are not immutable and can be influenced by situational or ecological variables reflected in the assessment design.
- The next generation must be prepared to cross disciplinary boundaries and move along the continuum of fundamental and applied work.

This essay's central take-home message is that assessment design, delivery, and test validity have changed significantly from 1900 to 1960 and more from 1960 to now, along with social, political, economic, cultural, scientific, and technological changes that have shaped our world. As such, the “over-the-shoulder look” back at some key moments in assessment set a course forward. Glancing at where we have been in test validity highlights the emergent meta-level trend toward explanation-focused thinking. Some scholars may argue that this was an emergent or unintentional trend because there is no recorded “meeting of the validity families,” in a manner of speaking, to carve up the assessment territory. I would suggest that the move to an explanation-focused view was concomitant with the evolution (or desire) for the development



of psychological science, as reflected, for example, in Cronbach and Meehl (1955).

As we took a retrospective look at the field of assessment while looking forward to the horizon for a glimpse of what lies in store, my offering to the field is the explanation-focused view of test validity, validation methods, and assessment research of which this essay presented a case for its need and a coherent description from this primitives and context in which it was developed and a detailed description of what it is as well as what I see as emerging on the horizon in terms of innovations in methods. The approach purposefully aims to push the boundaries of our validation practices, as Zumbo (2017) states, trending away from routine procedures toward an ecologically informed in vivo view of validation practices that are responsive to the cultural and social tectonic shifts of the last six decades highlighting how these social and cultural forces see concomitant changes in test validity in educational and psychological measurement.

### Acknowledgments

I am grateful to Lidia J. Jendzjowsky for her feedback and support while I wrote this paper. I also owe a debt of gratitude to my longtime collaborator, mentor, and friend, Donald W. Zimmerman, who, in 1986 took in a mathematical analyst looking for a new academic home. We collaborated for nearly the next 28 years on developing measure-theoretic mental test theory, some of which are reflected in the major themes of Sections 4.4 to 4.6 of this essay.

This research was undertaken, in part, thanks to funding in support of the Paragon UBC Professor of Psychometrics and Measurement, the Social Sciences and Humanities Research Council (SSHRC) of Canada, and the Canada Research Chairs Program in support of my Tier-1 Canada Research Chair in Psychometrics and Measurement.

### Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

### Orcid

Bruno D. Zumbo  <https://orcid.org/0000-0003-2885-5724>

### REFERENCES

- Addey, C., Maddox, B., & Zumbo, B.D. (2020) Assembled validity: Rethinking Kane's argument-based approach in the context of International Large-Scale Assessments (ILSAs), *Assessment in Education: Principles, Policy & Practice*, 27(6), 588-606. <https://doi.org/10.1080/0969594X.2020.1843136>
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. American Psychological Association.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME]. (1999). *Standards for educational and psychological testing*. American Educational Research Association.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. <https://www.testingstandards.net/open-access-files.html>
- American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51(2, Pt.2), 1-38. <https://doi.org/10.1037/h0053479>

- Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement*, 10, 67–78. <https://doi.org/10.1177/001316445001000105>
- Anastasi, A. (1954). *Psychological testing* (1<sup>st</sup> ed.). Macmillan.
- Angoff, W.H. (1988). Validity: An evolving concept. In: H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 19-32). Lawrence Erlbaum Associates.
- Bazire, M., & Brézillon, P. (2005). Understanding Context Before Using It. In: Dey, A., Kokinov, B., Leake, D., Turner, R. (eds) *modeling and using context. CONTEXT 2005. Lecture notes in computer science, vol. 3554*. Springer. [https://doi.org/10.1007/11508373\\_3](https://doi.org/10.1007/11508373_3)
- Bingham, W.V. (1937). *Aptitudes and aptitude testing*. Harper.
- Borsboom, D., Mellenbergh, G.J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061-1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Borsboom, D., Cramer, A.O.J., Kievit, R.A., Scholten, A.Z., & Frančić, S. (2009). The end of construct validity. In R.W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 135–170). IAP Information Age Publishing.
- Bronfenbrenner, U. (1979). *The ecology of human development*. Harvard University Press.
- Bronfenbrenner, U. (1994). Ecological models of human development. In T. Huston & T.N. Postlethwaith (Eds.), *International encyclopedia of education, 2nd ed., Vol. 3* (pp. 1643-1647). Elsevier Science.
- Buckingham, B.R. (1921). Intelligence and its measurement: A symposium. *Journal of Educational Psychology*, 12, 271–275.
- Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105. <https://doi.org/10.1037/h0046016>
- Carnap R. (1935). *Philosophy and logical syntax*. American Mathematical Society.
- Chen, M.Y., & Zumbo, B.D. (2017). Ecological framework of item responding as validity evidence: An application of multilevel DIF modeling using PISA data. In: Zumbo, B., Hubley, A. (eds) *Understanding and investigating response processes in validation research*. Springer, Cham. [https://doi.org/10.1007/978-3-319-56129-5\\_4](https://doi.org/10.1007/978-3-319-56129-5_4)
- ChoGlueck, C. (2018). The error is in the gap: Synthesizing accounts for societal values in science. *Philosophy of Science*, 85(4), 704-725. <https://doi.org/10.1086/699191>
- Clark, A. (1998). *Being there: Putting brain, body, and world together again*. MIT press.
- Clark, A. (2011). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford University Press.
- Courtis, S.A. (1921). Report of the standardization committee. *Journal of Educational Research*, 4(1), 78–90.
- Cronbach, L.J. (1971). Test validation. In: R.L. Thorndike (ed.) *Educational measurement, 2<sup>nd</sup> ed.* (pp. 443-507). American Council on Education.
- Cronbach, L.J. (1988). Five perspectives on the validity argument. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum Associates, Inc.
- Cronbach, L.J. (1989). Construct validation after thirty years. In R.L. Linn (ed.) *Intelligence: Measurement, theory, and public policy: Proceedings of a symposium in honor of Lloyd G. Humphreys* (pp. 147-171). University of Illinois Press.
- Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. <https://doi.org/10.1037/h0040957>
- Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511524059>
- de Ayala, R.J. (2009). [Review of Handbook of Statistics, Volume 26: Psychometrics, by C.R. Rao & S. Sinharay]. *Journal of the American Statistical Association*, 104(487), 1281–

1283. <http://www.jstor.org/stable/40592308>
- Dewey, J. (1938). *Logic: the theory of inquiry*. Holt.
- Douglas H. (2000) Inductive risk and values in science. *Philosophy of Science*, 67, 559–79. <https://doi.org/10.1086/392855>
- Douglas, H. (2003). The Moral Responsibilities of Scientists (Tensions between Autonomy and Responsibility). *American Philosophical Quarterly*, 40(1), 59-68. <http://www.jstor.org/stable/20010097>
- Douglas, H. (2004). The Irreducible Complexity of Objectivity. *Synthese* 138, 453–473. <https://doi.org/10.1023/B:SYNT.0000016451.18182.91>
- Douglas, H. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.
- Douglas, H. (2016), Values in science. In P. Humphries (ed.), *The Oxford Handbook of Philosophy of Science* (pp. 609-630). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199368815.013.28>
- Eid, M. (1996). Longitudinal confirmatory factor analysis for polytomous item responses: Model definition and model selection on the basis of stochastic measurement theory. *Methods of Psychological Research Online*, 1(4), 65-85.
- Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, 65, 241-261. <https://doi.org/10.1007/BF02294377>
- Elliott, K. (2011). *Is a little pollution good for you?: incorporating societal values in environmental research*. Oxford University Press.
- Embretson S.E. (Whitely). (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93(1), 179–197. <https://doi.org/10.1037/0033-2909.93.1.179>
- Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika*, 49(2), 175–186. <https://doi.org/10.1007/BF02294171>
- Embretson, S. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R.J., Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125– 150). Erlbaum.
- Embretson, S.E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3(3), 380-396. <https://doi.org/10.1037/1082-989X.3.3.380>
- Embretson, S.E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, 36(8), 449-455. <https://doi.org/10.3102/0013189X07311600>
- Embretson, S.E. (2016), Understanding Examinees' Responses to Items: Implications for Measurement. *Educational Measurement: Issues and Practice*, 35, 6-22. <https://doi.org/10.1111/emip.12117>
- Embretson, S., Schneider, L.M., & Roth, D.L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement*, 23, 13–32. <https://doi.org/10.1111/j.1745-3984.1986.tb00231.x>
- Fine, A.I. (1984). The natural ontological attitude (pp. 261-277). In J. Leplin (ed.), *Scientific realism*. University of California Press.
- Fox, J., Pychyl, T., & Zumbo, B.D. (1997). An investigation of background knowledge in the assessment of language proficiency. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma, (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 1996* (pp. 367 – 383). University of Jyväskylä Press.
- Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, 71(1), 5–19. <https://doi.org/10.2307/2024924>
- Galupo, M.P., Mitchell, R.C., & Davis, K.S. (2018). Face validity ratings of sexual orientation scales by sexual minority adults: Effects of sexual orientation and gender identity.

- Archives of Sexual Behavior*, 47(4), 1241–1250. <https://doi.org/10.1007/s10508-017-1037-y>
- Geiser, C., & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state-trait analyses. *Psychological Methods*, 17(2), 255–283. <https://doi.org/10.1037/a0026977>
- Giere, R.N. (1999). *Science without Laws*. University of Chicago Press.
- Giere, R.N. (2006). *Scientific perspectivism*. University of Chicago Press. <https://doi.org/10.7208/chicago/9780226292144.001.0001>
- Giere, R.N. (2010). *Explaining science: A cognitive approach*. University of Chicago Press.
- Gigerenzer, G., Swijtink, Z.G., Porter, T.M., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge University Press.
- Goffman, E. (1959). *The presentation of self in everyday life*. Doubleday.
- Goffman, E. (1964). The Neglected Situation. *American Anthropologist*, 66(6), 133–136. <http://www.jstor.org/stable/668167>
- Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, 33(2), 234–246. <https://doi.org/10.1111/j.2044-8317.1980.tb00610.x>
- Goldstein, H. (1994). Recontextualizing mental measurement. *Educational Measurement: Issues and Practice*, 12(1), 16–19, 43.
- Goldstein H. (1995). *Multilevel statistical models* (2<sup>nd</sup> edition). Edward Arnold/Halstead Press.
- Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, 42(2), 139–167. <https://doi.org/10.1111/j.2044-8317.1989.tb00905.x>
- Green, B. F. (1990). A comprehensive assessment of measurement. *Contemporary Psychology*, 35, 850–851.
- Green, C.D. (2015). Why psychology isn't unified, and probably never will be. *Review of General Psychology*, 19(3), 207–214. <https://doi.org/10.1037/gpr0000051>
- Guilford, J.P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6(4), 427–438. <https://doi.org/10.1177/001316444600600401>
- Guion, R.M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, 11(3), 385–398. <https://doi.org/10.1037/0735-7028.11.3.385>
- Gulliksen, H. (1950a). Intrinsic validity. *American Psychologist*, 5(10), 511–517. <https://doi.org/10.1037/h0054604>
- Gulliksen, H. (1950b). *Theory of mental tests*. John Wiley & Sons Inc. <https://doi.org/10.1037/13240-000>
- Gulliksen, H. (1961). Measurement of learning and mental abilities. *Psychometrika* 26, 93–107. <https://doi.org/10.1007/BF02289688>
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255–282. <https://doi.org/10.1007/BF02288892>
- Haig, B.D. (1999). Construct validation and clinical assessment. *Behaviour Change*, 16, 64–73.
- Haig, B.D. (2005a). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research*, 40(3), 303–329.
- Haig, B.D. (2005b). An abductive theory of scientific method. *Psychological Methods*, 10(4), 371–388. <https://doi.org/10.1037/1082-989X.10.4.371>
- Haig, B.D. (2009). Inference to the best explanation: A neglected approach to theory appraisal in psychology. *The American journal of psychology*, 122(2), 219–234.
- Haig, B.D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences*. MIT Press.
- Haig, B.D. (2018). Exploratory factor analysis, theory generation, and scientific method (pp.



- 65-88). In: *Method matters in psychology. Studies in applied philosophy, epistemology and rational ethics*, vol 45. Springer, Cham.
- Haig, B.D. (2019). The importance of scientific method for psychological science. *Psychology, Crime & Law*, 25(6), 527–541. <https://doi.org/10.1080/1068316X.2018.1557181>
- Haig, B.D. (in press). Repositioning construct validity theory: From nomological networks to pragmatic theories, and their evaluation by expiatory means. *Perspectives on Psychological Science*.
- Haig, B.D., & Evers, C.W. (2016). *Realist inquiry in social science*. Sage.
- Hattie, J., & Leeson, H. (2013). Future directions in assessment and testing in education and psychology. In K.F. Geisinger, B.A. Bracken, J.F. Carlson, J.-I. C. Hansen, N.R. Kuncel, S.P. Reise, & M.C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, vol. 3. testing and assessment in school psychology and education* (pp. 591–622). American Psychological Association. <https://doi.org/10.1037/14049-028>
- Hempel, C.G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. The Free Press.
- Hicks, D.J. (2014). A new direction for science and values. *Synthese*, 191(14), 3271–3295. <http://www.jstor.org/stable/24026188>
- Higgins, N.C., Zumbo, B.D., & Hay, J.L. (1999). Construct validity of attributional style: Modeling context-dependent item sets in the attributional style questionnaire. *Educational and Psychological Measurement*, 59(5), 804-820. <https://doi.org/10.1177/0131649921970152>
- Holman, B., & Wilholt, T. (2022). The new demarcation problem. *Studies in history and philosophy of science*, 91, 211-220. <https://doi.org/10.1016/j.shpsa.2021.11.011>
- Hubley, A.M., & Zumbo, B.D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology*, 123(3), 207-215. <https://doi.org/10.1080/00221309.1996.9921273>
- Hubley, A.M., & Zumbo, B.D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, 103(2), 219–230. <https://doi.org/10.1007/s11205-011-9843-4>
- Hubley, A.M., & Zumbo, B.D. (2013). Psychometric characteristics of assessment procedures: An overview. In Kurt F. Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology, 1* (pp. 3-19). American Psychological Association Press. <https://doi.org/10.1037/14047-001>
- Hubley, A.M., & Zumbo, B.D. (2017). Response processes in the context of validity: Setting the stage. In B.D. Zumbo & A.M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 1–12). Springer International Publishing/Springer Nature. [https://doi.org/10.1007/978-3-319-56129-5\\_1](https://doi.org/10.1007/978-3-319-56129-5_1)
- Hull, C.L. (1935). The conflicting psychologies of learning: A way out. *Psychological Review*, 42(6), 491–516. <https://doi.org/10.1037/h0058665>
- Jonson, J.L., & Plake, B.S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement*, 58(5), 736-753. <https://doi.org/10.1177/0013164498058005002>
- Kaldis, B. (2013). Kinds: natural kinds versus human kinds. In *Encyclopedia of Philosophy and the Social Sciences, 2*, (pp. 515-518). SAGE Publications, Inc. <https://doi.org/10.4135/9781452276052>
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527–535. <https://doi.org/10.1037/0033-2909.112.3.527>
- Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319-342. <https://doi.org/10.1111/j.1745-3984.2001.tb01130.x>
- Kane, M. (2004). Certification testing as an illustration of argument-based validation.



- Measurement: Interdisciplinary Research and Perspective*, 2(3), 135-170. [https://doi.org/10.1207/s15366359mea0203\\_1](https://doi.org/10.1207/s15366359mea0203_1)
- Kane, M. (2006). Validation. In R. Brennan (Ed.) *Educational measurement* (4th ed., pp. 17-64). American Council on Education and Praeger.
- Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, 29(1), 3-17. <https://doi.org/10.1177/0265532211417210>
- Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1-73. <https://doi.org/10.1111/jedm.12000>
- Kane, M. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23(2), 198–211. <https://doi.org/10.1080/0969594X.2015.1060192>
- Kincaid, H. (2000). Global arguments and local realism about the social sciences. *Philosophy of Science*, 67(S3), S667-S678. <https://doi.org/10.1086/392854>
- Koch, T., Eid, M., & Lochner, K. (2018). Multitrait-multimethod-analysis: The psychometric foundation of CFA-MTMM models. In P. Irwing, T. Booth, & D.J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 781-846). Wiley Blackwell. <https://doi.org/10.1002/9781118489772.ch25>
- Koch, T., Schultze, M., Eid, M., & Geiser, C. (2014). A longitudinal multilevel CFA-MTMM model for interchangeable and structurally different methods. *Frontiers in Psychology*, 5, Article 311. <https://doi.org/10.3389/fpsyg.2014.00311>
- Kroc, E., & Zumbo, B.D. (2018). Calibration of measurements. *Journal of Modern Applied Statistical Methods*, 17(2), eP2780. <https://digitalcommons.wayne.edu/jmasm/vol17/iss2/17/>
- Kroc, E., & Zumbo, B.D. (2020). A transdisciplinary view of measurement error models and the variations of  $X = T + E$ . *Journal of Mathematical Psychology*, 98, 102372. <https://doi.org/10.1016/j.jmp.2020.102372>
- Kuhn, T.S. (1962). *The structure of scientific revolutions*. University of Chicago Press.
- Kuhn, T.S. (1970). *The structure of scientific revolutions* (2<sup>nd</sup> ed.). University of Chicago Press.
- Kuhn, T.S. (1977). *The essential tension: Selected studies in scientific tradition and change*. University of Chicago Press.
- Kuhn, T.S. (1996). *The structure of scientific revolutions* (3<sup>rd</sup> ed.). University of Chicago Press.
- Lakatos I. (1976). *Falsification and the methodology of scientific research programmes. Can theories be refuted?* (pp. 205–259). Springer.
- Lane, S., Zumbo, B.D., Abedi, J., Benson, J., Dossey, J., Elliott, S.N., Kane, M., Linn, R., Paredes-Ziker, C., Rodriguez, M., Schraw, G., Slattery, J., Thomas, V., & Willhoft, J. (2009). Prologue: An Introduction to the Evaluation of NAEP. *Applied Measurement in Education*, 22(4), 309-316. <https://doi.org/10.1080/08957340903221436>
- Lennon, R.T. (1956). Assumptions Underlying the Use of Content Validity. *Educational and Psychological Measurement*, 16(3), 294-304. <https://doi.org/10.1177/001316445601600303>
- Lewis, C. (1986). Test theory and psychometrika: The past twenty-five years. *Psychometrika*, 51(1), 11–22. <https://doi.org/10.1007/BF02293995>
- Li, Z., & Zumbo, B.D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicológica*, 30(2), 343–370. <https://www.uv.es/psicologica/articulos2.09/11LI.pdf>
- Lipton, P. (2004). *Inference to the best explanation* (2<sup>nd</sup> ed.). Routledge. <https://doi.org/10.4324/9780203470855>
- Lissitz, R.W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437–448. <https://doi.org/10.3102/0013189X07311286>

- Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635-694 (Monograph Supp. 9).
- Maddox, B. (2015). The neglected situation: assessment performance and interaction in context. *Assessment in Education: Principles, Policy & Practice*, 22(4), 427-443. <https://doi.org/10.1080/0969594X.2015.1026246>
- Maddox, B., Zumbo, B.D. (2017). Observing testing situations: Validation as Jazz. In: B.D. Zumbo, A.M. Hubley (eds) *Understanding and investigating response processes in validation research*. Springer, Cham. [https://doi.org/10.1007/978-3-319-56129-5\\_10](https://doi.org/10.1007/978-3-319-56129-5_10)
- Maddox, B., Zumbo, B.D., Tay-Lim, B. S.-H., & Demin Qu, I. (2015). An anthropologist among the psychometricians: Assessment events, ethnography and DIF in the Mongolian Gobi. *International Journal of Testing*, 15(4), 291-309. <https://doi.org/10.1080/15305058.2015.1017103>
- Markus, K.A. (1998). Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible?. *Social Indicators Research*, 45, 7-34. <https://doi.org/10.1023/A:1006960823277>
- MacCorquodale, K., & Meehl, P.E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 55(2), 95-107. <https://doi.org/10.1037/h0056029>
- Mehrens, W.A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16(2), 16-18.
- Messick, S. (1972). Beyond structure: In search of functional models of psychological process. *Psychometrika*, 37(4, Pt. 1), 357-375. <https://doi.org/10.1007/BF02291215>
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955-966.
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35, 1012-1027.
- Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In: H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 33-45). Lawrence Erlbaum Associates.
- Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3<sup>rd</sup> ed., pp. 13-103). Macmillan.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Messick, S. (1998). Test validity: A matter of consequence [Special issue]. *Social Indicators Research*, 45, 35-44. <https://doi.org/10.1023/A:1006964925094>
- Messick, S. (2000). Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. In: Goffin, R.D., Helmes, E. (eds) *Problems and solutions in human assessment*. Springer. [https://doi.org/10.1007/978-1-4615-4397-8\\_1](https://doi.org/10.1007/978-1-4615-4397-8_1)
- Millman, J. (1979). Reliability and validity of criterion-referenced test scores. In: R. Traub (Ed.), *New directions for testing and measurement: Methodological developments*. Jossey-Bass.
- Mosier, C.I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement*, 7(2), 191-205. <https://doi.org/10.1177/001316444700700201>
- Nickles, T. (2017). Cognitive illusions and nonrealism: Objections and replies. In: Agazzi, E. (eds) *Varieties of Scientific Realism: Objectivity and truth in science* (pp. 151-163). Springer, Cham. [https://doi.org/10.1007/978-3-319-51608-0\\_8](https://doi.org/10.1007/978-3-319-51608-0_8)
- Novick, M.R. (1966). The axioms and principal results of classical test theory. *Journal of*

- Mathematical Psychology*, 3(1), 1–18. [https://doi.org/10.1016/0022-2496\(66\)90002-2](https://doi.org/10.1016/0022-2496(66)90002-2)
- O’Leary, T.M., Hattie, J.A.C., & Griffin, P. (2017). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues and Practice*, 36, 16-23. <https://doi.org/10.1111/emip.12141>
- Padilla, J.L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, 26, 136–144. <https://doi.org/10.7334/psicothema2013.259>
- Padilla, J.L., & Benítez, I. (2017). A rationale for and demonstration of the use of DIF and mixed methods. In: Zumbo, B.D., Hubley, A.M. (eds) *Understanding and investigating response processes in validation research* (pp. 193–210). Springer, Cham. [https://doi.org/10.1007/978-3-319-56129-5\\_1](https://doi.org/10.1007/978-3-319-56129-5_1)
- Pellicano, E., & den Houting, J. (2022). Annual research review: Shifting from “normal science” to neurodiversity in autism science. *Journal of Child Psychology and Psychiatry*, 63, 381–396. <https://doi.org/10.1111/jcpp.13534>
- Persson, J., & Ylikoski, P. (Eds.). (2007). *Rethinking explanation* (Boston Studies in the Philosophy of Science, Vol. 252). Springer.
- Pitt, J.C. (Ed.) (1988). *Theories of explanation*. Oxford University Press.
- Popham, W.J. (1997). Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice*, 16(2), 9-13.
- Psillos, S. (2022). Realism and theory change in science. In: Zalta, E.N., Nodelman, U. (eds.) *The Stanford encyclopedia of philosophy*. <https://plato.stanford.edu/archives/fall2022/entries/realism-theory-change/>
- Rao, C.R., & Sinharay, S. (Eds.). (2007). *Handbook of statistics, Volume 26: Psychometrics*. Elsevier.
- Raykov, T. (1992), On structural models for analyzing change. *Scandinavian Journal of Psychology*, 33, 247-265. <https://doi.org/10.1111/j.1467-9450.1992.tb00914.x>
- Raykov, T. (1998a). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement*, 22(4), 375-385. <https://doi.org/10.1177/014662169802200407>
- Raykov, T. (1998b). A method for obtaining standard errors and confidence intervals of composite reliability for congeneric items. *Applied Psychological Measurement*, 22(4), 369-374. <https://doi.org/10.1177/014662169802200406>
- Raykov, T. (1999). Are simple change scores obsolete? An approach to studying correlates and predictors of change. *Applied Psychological Measurement*, 23(2), 120-126. <https://doi.org/10.1177/01466219922031248>
- Raykov, T. (2001), Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology*, 54, 315-323. <https://doi.org/10.1348/000711001159582>
- Raykov, T., & Marcoulides, G.A. (2011). *Introduction to psychometric theory*. Routledge.
- Raykov, T., & Marcoulides, G.A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and Psychological Measurement*, 76(2), 325–338. <https://doi.org/10.1177/0013164415576958>
- Reichenbach H. (1977). Philosophie der Raum-Zeit-Lehre. In: Kamlah, A., Reichenbach, M. (eds) *Philosophie der Raum-Zeit-Lehre. Hans Reichenbach, vol 2*. Vieweg+Teubner Verlag, Wiesbaden.
- Roberts, B.W. (2007). Contextualizing personality psychology. *Journal of Personality*, 75(6), 1071–1082. <https://doi.org/10.1111/j.1467-6494.2007.00467.x>
- Rome, L., & Zhang, B. (2018). Investigating the effects of differential item functioning on proficiency classification. *Applied psychological measurement*, 42(4), 259–274. <https://doi.org/10.1177/0146621617726789>
- Rozeboom, W.W. (1966). *Foundations of the theory of prediction*. Dorsey.

- Rulon, P.J. (1946). On the validity of educational tests. *Harvard Educational Review*, 16, 290-296.
- Salmon, W. (1990). *Four decades of scientific explanation*. University of Minnesota Press.
- Schaffner, K.F. (2020). A comparison of two neurobiological models of fear and anxiety: A “construct validity” application? *Perspectives on Psychological Science*, 15(5), 1214-1227. <https://doi.org/10.1177/1745691620920860>
- Schaffner, K.F. (1993). *Discovery and explanation in biology and medicine*. University of Chicago Press.
- Searle, J.R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press.
- Searle, J.R. (1979). *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511609213>
- Sells, S.B. (ed.) (1963). *Stimulus determinants of behavior*. Ronald Press.
- Shear, B.R., Zumbo, B.D. (2014). What counts as evidence: A review of validity studies in educational and psychological measurement. In: Zumbo, B.D., Chan, E.K.H. (eds) *Validity and validation in social, behavioral, and health sciences* (pp. 91-111). Springer, Cham. [https://doi.org/10.1007/978-3-319-07794-9\\_6](https://doi.org/10.1007/978-3-319-07794-9_6)
- Shepard, L.A. (1993). Evaluating test validity. *Review of Research in Education*, 19(1), 405-450. <https://doi.org/10.3102/0091732X019001405>
- Shepard, L.A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16, 5-8, 13, 24.
- Sinnott-Armstrong, W., & Fogelin, R.J. (2010). *Understanding arguments: An introduction to informal logic*. Wadsworth Cengage Learning.
- Sireci, S.G. (1998). The construct of content validity [Special issue]. *Social Indicators Research* 45, 83–117. <https://doi.org/10.1023/A:1006985528729>
- Sireci, S.G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R.W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19–37). IAP Information Age Publishing.
- Sireci, S.G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, 50, 99-104. <https://doi.org/10.1111/jedm.12005>
- Sireci, S.G. (2020). De-“constructing” test validation. *Chinese/English Journal of Educational Measurement and Evaluation*, 1(1), Article 3. <https://www.ce-jeme.org/journal/vol1/iss1/3>
- Slaney, K.L., & Racine, T.P. (2013). What’s in a name? Psychology’s ever evasive construct. *New Ideas in Psychology*, 31(1), 4-12. <https://doi.org/10.1016/j.newideapsych.2011.02.003>
- Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), 72–101. <https://doi.org/10.2307/1412159>
- Steyer, R. (1988). Conditional expectations: An introduction to the concept and its applications in empirical sciences. *Methodika*, 2, 53-78.
- Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika*, 3, 25-60.
- Steyer, R., Ferring, D., & Schmitt, M.J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, 8(2), 79–98.
- Steyer, R., Majcen, A.-M., Schwenkmezger, P., & Buchner, A. (1989). A latent state-trait anxiety model and its application to determine consistency and specificity coefficients. *Anxiety Research*, 1(4), 281–299. <https://doi.org/10.1080/08917778908248726>
- Steyer, R., & Schmitt, M. (1990). Latent state-trait models in attitude research. *Quality & Quantity*, 24, 427–445. <https://doi.org/10.1007/BF00152014>



- Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state–trait theory and research in personality and individual differences. *European Journal of Personality*, 13(5), 389-408. [https://doi.org/10.1002/\(SICI\)1099-0984\(199909/10\)13:5<389::AID-PER361>3.0.CO;2-A](https://doi.org/10.1002/(SICI)1099-0984(199909/10)13:5<389::AID-PER361>3.0.CO;2-A)
- Stone, J., & Zumbo, B.D. (2016). Validity as a pragmatist project: A global concern with local application. In: Aryadoust V., & Fox J. (eds.) *Trends in language assessment research and practice* (pp. 555–573). Cambridge Scholars Publishing.
- Suppes, P. (1969). Models of data. In: *Studies in the methodology and foundations of science. Synthese Library, vol 22*. Springer. [https://doi.org/10.1007/978-94-017-3173-7\\_2](https://doi.org/10.1007/978-94-017-3173-7_2)
- Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12(3), 435-467. <https://doi.org/10.1017/S0140525X00057046>
- Thagard, P. (1992). *Conceptual revolutions*. Princeton University Press. <http://www.jstor.org/stable/j.ctv36zq4g>
- Tolman, C.W. (1991). Review of constructing the subject: Historical origins of psychological research [Review of the book *Constructing the subject: Historical origins of psychological research*, by K. Danziger]. *Canadian Psychology*, 32(4), 650–652. <https://doi.org/10.1037/h0084651>
- Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.
- van Fraassen, B.C. (1980). *The scientific image*. Oxford University Press. <https://doi.org/10.1093/0198244274.001.0001>
- van Fraassen, B.C. (1985). Empiricism in the philosophy of science. In: Churchland P.M., & Hooker C.A. (eds.) *Images of science: Essays on realism and empiricism* (pp. 245-308). University of Chicago Press.
- van Fraassen, B.C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford University Press.
- van Fraassen, B.C. (2012). Modeling and measurement: The criterion of empirical grounding. *Philosophy of Science*, 79(5), 773–784. <https://doi.org/10.1086/667847>
- Varela, F.J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. The MIT Press. <https://doi.org/10.7551/mitpress/6730.001.0001>
- Wallin, A. (2007). Explanation and environment. In: Persson, J., Ylikoski, P. (eds) *Rethinking explanation. Boston studies in the philosophy of science, (pp. 163-175), vol 252*. Springer, Dordrecht. [https://doi.org/10.1007/978-1-4020-5581-2\\_12](https://doi.org/10.1007/978-1-4020-5581-2_12)
- Wapner, S., & Demick, J. (2002). The increasing contexts of context in the study of environment behavior relations. In R.B. Bechtel & A. Churchman (eds.) *Handbook of environmental psychology* (pp. 3–14). John Wiley & Sons, Inc.
- Watson, J.B. (1913). Psychology as the behaviorist views it. *Psychological Review*, 20(2), 158–177. <https://doi.org/10.1037/h0074428>
- Whitely (Embretson), S.E. (1977). Information-processing on intelligence test items: Some response components. *Applied Psychological Measurement*, 1, 465-476. <https://doi.org/10.1177/014662167700100402>
- Wiley, D.E. (1991). Test validity and invalidity reconsidered. In: R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: a volume in honor of Lee J. Cronbach* (pp. 75-107). Erlbaum.
- Woitschach, P., Zumbo, B.D., & Fernández-Alonso, R. (2019). An ecological view of measurement: Focus on multilevel model explanation of differential item functioning. *Psicothema*, 31(2), 194–203. <https://doi.org/10.7334/psicothema2018.303>
- Woodward, J. (1989). Data and phenomena. *Synthese*, 79, 393-472. <https://doi.org/10.1007/BF00869282>
- Wu, A.D., & Zumbo, B.D. (2008). Understanding and using mediators and moderators. *Social Indicators Research*, 87, 367–392. <https://doi.org/10.1007/s11205-007-9143-1>



- Wu, A.D., Zumbo, B.D., & Marshall, S.K. (2014). A method to aid in the interpretation of EFA results: An application of Pratt's measures. *International Journal of Behavioral Development, 38*(1), 98-110. <https://doi.org/10.1177/0165025413506143>
- Yang, Y., Read, S.J., & Miller, L.C. (2009). The concept of situations. *Social and Personality Psychology Compass, 3*(6), 1018-1037. <https://doi.org/10.1111/j.1751-9004.2009.00236.x>
- Zimmerman, D.W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika, 40*(3), 395-412. <https://doi.org/10.1007/BF02291765>
- Zimmerman, D.W., & Zumbo, B.D. (2001). The geometry of probability, statistics, and test theory. *International Journal of Testing, 1*(3-4), 283-303. <https://doi.org/10.1080/15305058.2001.9669476>
- Zumbo, B.D. (Ed.). (1998). *Validity theory and the methods used in validation: perspectives from the social and behavioral sciences*. In: Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, [Special volume], Vol. 45, Issues 1-3. Springer International Publishing.
- Zumbo, B.D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. *Advances in social science methodology, 5*(1), 269-304.
- Zumbo, B.D. (2005, July). *Reflections on validity at the intersection of psychometrics, scaling, philosophy of inquiry, and language testing* [Samuel J. Messick Memorial Award Lecture]. LTRC, the 27th Language Testing Research Colloquium, Ottawa, Canada.
- Zumbo, B.D. (2007a). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 45-79). Elsevier.
- Zumbo, B.D. (2007b). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly, 4*(2), 223-233. <https://doi.org/10.1080/15434300701375832>
- Zumbo, B.D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R.W. Lissitz (ed.) *The concept of validity: Revisions, new directions, and applications* (pp. 65-82). IAP Information Age Publishing.
- Zumbo, B.D. (2010, September). *Measurement validity and validation: A meditation on where we have come from and the state of the art today* [Invited address]. Presented at the International conference on outcomes measurement, US National Institutes of Health, Bethesda, MD.
- Zumbo, B.D. (2015, November). *Consequences, side effects and the ecology of testing: Keys to considering assessment "in vivo"* [Plenary address]. Annual Meeting of the Association for Educational Assessment – Europe (AEA Europe), Glasgow, Scotland. <https://youtu.be/0L6Lr2BzuSQ>
- Zumbo, B.D. (2016). *Standard Setting Methodology* [Invited address]. "Applied Physiology Physical Employment Standards - Current Issues and Challenges" at the Canadian Society for Exercise Physiology (CSEP) conference, Victoria, Canada.
- Zumbo, B.D. (2017). Trending away from routine procedures, toward an ecologically informed in vivo view of validation practices. *Measurement: Interdisciplinary Research and Perspectives, 15*(3-4), 137-139. <https://doi.org/10.1080/15366367.2017.1404367>
- Zumbo, B.D. (2018a, April). *Methodologies used to ensure fairness and equity in the assessment of students' educational outcomes* [Invited presentation and panel session]. AERA Presidential Symposium "Methodology and equity: An international perspective" at the Annual Meeting of the American Educational Research Association (AERA), New York, NY.
- Zumbo, B.D. (2018b, July). *The reports of DIF's death are greatly exaggerated; It is like a Phoenix rising from the ashes* [Keynote Address]. The 11th Conference of the International Test Commission, Montreal, Canada.

- Zumbo, B.D. (2019). Foreword: Tensions, Intersectionality, and What Is on the Horizon for International Large-Scale Assessments in Education. In B. Maddox (Ed.), *International large-scale assessments in education: Insider research perspectives* (pp. xii–xiv). Bloomsbury Publishing. <https://doi.org/10.5040/9781350023635>
- Zumbo, B.D. (2021). *A novel multimethod approach to investigate whether tests delivered at a test centre are concordant with those delivered remotely online* [Research Monograph]. UBC Psychometric Research Series, University of British Columbia. <http://dx.doi.org/10.14288/1.0400581>
- Zumbo, B.D. (2023a). *Validity theories, frameworks and practices in using tests and measures: an over-the-shoulder look back at validity while also looking to the horizon* [Invited Address]. Ciclo Formazione Metodologica (FORME), Dipartimento di Psicologia, Università Cattolica Del Sacro Cuore. [https://brunozumbo.com/?page\\_id=31](https://brunozumbo.com/?page_id=31)
- Zumbo, B.D. (2023b). *Test validation and Bayesian statistical frameworks to estimate the magnitude and corresponding uncertainty of washback effects of test preparation* [Research Monograph]. UBC Psychometric Research Series, University of British Columbia. <https://dx.doi.org/10.14288/1.0435197>
- Zumbo, B.D. (2023c, October). *The Challenges and Promise of Embracing the Many Ways of Being Human: Toward an Ecologically Informed In Vivo View of Validation Practices* [Invited Address]. Symposium on Inclusive Educational Assessment, Neurodiversity and Disability. Hughes Hall, University of Cambridge.
- Zumbo, B.D., & Chan, E.K.H. (Eds.). (2014a). *Validity and validation in social, behavioral, and health sciences*. Springer International Publishing/Springer Nature. <https://doi.org/10.1007/978-3-319-07794-9>
- Zumbo, B.D., & Chan, E.K.H. (2014b). Reflections on validation practices in the social, behavioral, and health sciences. In: Zumbo, B.D., Chan, E.K.H. (eds) *Validity and validation in social, behavioral, and health sciences* (pp. 321-327). Springer, Cham. [https://doi.org/10.1007/978-3-319-07794-9\\_19](https://doi.org/10.1007/978-3-319-07794-9_19)
- Zumbo, B.D., & Chan, E.K.H. (2014c). Setting the stage for validity and validation in social, behavioral, and health sciences: Trends in validation practices. In: Zumbo, B.D., Chan, E.K.H. (eds) *Validity and validation in social, behavioral, and health sciences* (pp. 3-8). Springer, Cham. [https://doi.org/10.1007/978-3-319-07794-9\\_1](https://doi.org/10.1007/978-3-319-07794-9_1)
- Zumbo, B.D., & Forer, B. (2011). Testing and measurement from a multilevel view: Psychometrics and validation. In J.A. Bovaird, K.F. Geisinger, & C.W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K–12 settings* (pp. 177–190). American Psychological Association. <https://doi.org/10.1037/12330-011>
- Zumbo, B.D., & Gelin, M.N. (2005). A matter of test bias in educational policy research: bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies*, 5, 1–23. URL: <https://files.eric.ed.gov/fulltext/EJ846827.pdf>
- Zumbo, B. D., & Hubley, A. M. (2016). Bringing consequences and side effects of testing and assessment to the foreground. *Assessment in Education: Principles, Policy & Practice*, 23(2), 299–303. <https://doi.org/10.1080/0969594X.2016.1141169>
- Zumbo, B.D., & Hubley, A.M. (Eds.). (2017). *Understanding and investigating response processes in validation research*. Springer International Publishing/Springer Nature. <https://doi.org/10.1007/978-3-319-56129-5>
- Zumbo, B.D., & Kroc, E. (2019). A Measurement Is a Choice and Stevens’ scales of measurement do not help make it: A response to chalmers. *Educational and Psychological Measurement*, 79(6), 1184-1197. <https://doi.org/10.1177/0013164419844305>
- Zumbo, B.D., Liu, Y., Wu, A.D., Forer, B., Shear, B.R. (2017). National and international

- 
- educational achievement testing: A case of multi-level validation framed by the ecological model of item responding. In B.D. Zumbo & A.M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 341-362). Springer International Publishing/Springer Nature. [https://doi.org/10.1007/978-3-319-56129-5\\_18](https://doi.org/10.1007/978-3-319-56129-5_18)
- Zumbo, B.D., Liu, Y., Wu, A.D., Shear, B.R., Olvera Astivia, O.L., & Ark, T.K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, *12*(1), 136-151. <https://doi.org/10.1080/15434303.2014.972559>
- Zumbo, B.D., Maddox, B., & Care, N.M. (2023). Process and product in computer-based assessments: Clearing the ground for a holistic validity framework. *European Journal of Psychological Assessment*, *39*(4), 252–262. <https://doi.org/10.1027/1015-5759/a000748>
- Zumbo, B.D., & Padilla, J.-L. (2020). The interplay between survey research and psychometrics, with a focus on validity theory. In P.C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G.B. Willis, & A. Wilmot (Eds.), *Advances in questionnaire design, development, evaluation and testing* (pp. 593-612). John Wiley & Sons, Inc.. <https://doi.org/10.1002/9781119263685.ch24>
- Zumbo, B.D., Pychyl, T.A., & Fox, J.A. (1993). Psychometric properties of the CAEL assessment, II: An examination of the dependability/reliability of placement decisions. *Carleton Papers in Applied Language Studies*, *10*, 13-27.
- Zumbo, B.D., & Rupp, A.A. (2004). Responsible modeling of measurement data for appropriate inferences: important advances in reliability and validity theory. In David Kaplan (ed.) *The SAGE handbook of quantitative methodology for the social sciences* (pp. 74-93). SAGE Publications, Inc. <https://doi.org/10.4135/9781412986311>
- Zumbo, B.D., & Shear, B.R. (2011, October). *The concept of validity and some novel validation methods* [Lecture/Workshop, half-day]. The 42nd annual Northeastern Educational Research Association (NERA) meeting, Rocky Hill, CT.