e-ISSN: 2148-7456 | a peer-reviewed online journal | hosted by DergiPark

International Journal of Assessment Tools in Education

Volume: 10 Issue: Special Issue

December 2023

https://dergipark.org.tr/en/pub/ijate

e-ISSN: 2148-7456

Volume: 10 Issue: Special Issue 2023

Special Issue Editor Dr. Okan BULUT
Address University of Alberta, Edmonton, Canada
E-mail bulut@ualberta.ca
ijate.editor@gmail.com

Journal Contact Dr. Eren Can AYBEK
Address Department of Educational Sciences, Pamukkale University, Faculty of Education, Kinikli Yerleskesi, Denizli, 20070, Türkiye
Phone +90 258 296 31050
Fax +90 258 296 1200
E-mail erencanaybek@gmail.com

Address Dr. Anil KANDEMIR
Department of Educational Sciences, Agri Ibrahim Cecen University, Faculty of Education, Agri, Türkiye
akandemir@agri.edu.tr

Publisher Dr. Izzet KARA
Address Pamukkale University, Education Faculty, Kinikli Campus, 20070 Denizli, Türkiye
Phone +90 258 296 1036
Fax +90 258 296 1200
E-mail ikara@pau.edu.tr
ijate.editor@gmail.com

Frequency 4 issues per year (March, June, September, December)
Online ISSN 2148-7456
Website https://dergipark.org.tr/en/pub/ijate
https://ijate.net/index.php/ijate

Cover Design Merve SENTURK

International Journal of Assessment Tools in Education

International Journal of Assessment Tools in Education (IJATE) accepts original research on the design, analysis and use of evaluation along with assessment to enhance comprehension of the performance and quality of stakeholders in educational settings. IJATE is pleased to receive discriminating theoretical and empirical manuscripts (quantitative or qualitative) which could direct significant national and international argumentations in educational policy and practice.

IJATE, as an online journal, is hosted by DergiPark [TUBITAK-ULAKBIM (The Scientific and Technological Research Council of Türkiye)].

In IJATE, there are no charges under any procedure for submitting or publishing an article.

Indexes and Platforms:

- Emerging Sources Citation Index (ESCI)
- Education Resources Information Center (ERIC)
- TR Index (ULAKBIM),
- EBSCOhost,
- SOBIAD,
- JournalTOCs,
- MIAR (Information Matrix for Analysis of the Journals),
- idealonline,
- CrossRef,
- ResearchBib,
- Index Copernicus International

From Special Issue Editor:

Dear Readers,

In the ever-evolving realm of education, the landscape of measurement and evaluation has undergone a profound metamorphosis, primarily driven by the relentless advance of digital assessment tools. These tools, ranging from computer-based tests to automated essay scoring systems, have not only revolutionized how we gauge learning outcomes but have also opened new avenues for educators to tailor their approaches to the unique needs of each learner. As we stand at the nexus of technological innovation and educational practice, it is imperative to reflect on the past, acknowledge the present, and envision the future of educational assessment.

This special issue on "*Educational Measurement and Evaluation: Lessons from the Past, Visions for the Future*," presented by the International Journal of Assessment Tools in Education (IJATE), marks a significant milestone—the 100th anniversary of the Republic of Türkiye. In celebrating this historic occasion, we have curated a collection of ten articles that delve into various facets of educational measurement and evaluation. Each article serves as a testament to the remarkable journey we have undertaken in understanding and enhancing the educational assessment landscape. For this special issue, we specifically asked researchers to discuss the evolution of educational measurement and evaluation concepts and their vision for the future of these concepts. All articles submitted for publication went through a rigorous peer review based on the review standards established by IJATE.

Within the inaugural article of this issue, Zumbo delivers an outstanding exploration of test validity, imparting his insights on methodological innovations in explanation-focused validity. This paper is poised to be a cornerstone resource, equipping both researchers and practitioners with a nuanced understanding of test validity spanning historical roots to contemporary perspectives. Concurrently, the sixth article, penned by Mor and Karatoprak Ersen in this issue, delves into the realm of test validity. The authors meticulously examine the ramifications of prevailing validity frameworks within the context of classroom assessment, enriching the discourse on the crucial subject of test validity.

The second article by Arici and Kutlu (in this issue) focuses on factors related to collaborative problem-solving skills. Leveraging the outcomes of PISA 2015, the authors meticulously scrutinize both direct and indirect factors influencing the collaborative problem-solving aptitude of students in Türkiye. Simultaneously, the eighth article, contributed by He in this issue, also centers on collaborative problem-solving skills within large-scale assessments. The author evaluates the intricacies of item design and scoring and engages in a thoughtful discussion on potential approaches to gauge students' proficiency more accurately in collaborative problem-solving. Collectively, these articles contribute to the evolving discourse surrounding collaborative problem-solving, offering valuable insights for educators, researchers, and policymakers.

In the third article, Schwarz et al. (in this issue) describe a data pipeline for digital large-scale assessments (the authors refer to this as "e-large-scale assessments"). Employing the versatile R programming language, the authors skillfully showcase the automation of various data analysis steps. These include data transformation, psychometric analyses grounded in Classical Test Theory and Item Response Theory, and the streamlined generation of score reports. The article provides a comprehensive demonstration of how a meticulously designed data pipeline can automate these crucial processes, offering an insightful guide for practitioners and researchers engaged in large-scale assessments.

The fourth and fifth articles center on the automation of two pivotal psychometric tasks—item development and scoring—through the application of advanced computer algorithms. In their contribution to this issue, Sayin et al. showcase the efficacy of template-based automatic item generation in crafting non-verbal items for a visual reasoning test. The study's findings underscore the potential of automatic item generation in expeditiously building a substantial

repository of items. Similarly, Firoozi et al. delve into automated essay scoring in the fifth article, elucidating the application of large-language models. The authors intricately describe and illustrate the framework of automated essay scoring systems within the specific context of the Turkish language. Collectively, these articles illuminate the transformative impact of advanced algorithms in revolutionizing key psychometric processes, offering valuable insights into the future of assessment methodologies.

Taskin Bedizel (in this issue) unfolds the outcomes of a comprehensive bibliometric analysis, spanning publications from 1994 to September 2023, focused on the intersection of artificial intelligence and educational assessment research. The insights derived from this study not only offer a panoramic view of the evolution over time but also provide valuable guidance for researchers and practitioners navigating the trajectory of AI-powered assessment tools in education.

The seventh and tenth articles delve into the analysis of process data derived from international large-scale assessments. In this issue, Yilmaz Kogar and Soysal leverage item response time to investigate the impact of factors like item difficulty, content, and cognitive domain on problem-solving duration in TIMSS 2019. The authors draw upon samples from 4th-grade students in Singapore and Turkey, conducting a comparative analysis of aberrant response behaviors, including rapid guessing, between the two nations. In a parallel exploration, Ersan and Parlak utilize TIMSS 2019 data to scrutinize the influence of on-screen calculators on students' performance in the Problem Solving and Inquiry tasks. Their findings reveal a positive association between the use of on-screen calculators and the likelihood of correct item responses, highlighting the value of process data in understanding students' response behaviors in digital assessments.

I hope that the articles featured in this special issue will inspire further dialogue, spark new ideas, and contribute to the ongoing evolution of educational assessment as we embark on the next century of progress. As we navigate the complexities of the digital age, the featured articles will offer insights into the ways technology has reshaped our approaches to educational assessment. From exploring validity frameworks that underpin assessment methodologies to investigating the intricacies of automated essay scoring, automatic item generation, collaborative problem-solving, and digital large-scale assessments, these contributions encapsulate the breadth of advancements that have shaped the past and continue to mold the future of educational measurement. The lessons gleaned from the past and the visions articulated for the future converge in this special issue, offering a comprehensive exploration of the multifaceted world of educational measurement and evaluation.

In drawing this editorial summary to a close, I wish to express my heartfelt gratitude to the esteemed authors whose expertise and dedication have profoundly enriched this special issue. A special acknowledgment extends to all the diligent reviewers who generously shared their valuable insights, contributing significantly to the refinement of the submitted papers. Additionally, my sincere appreciation goes to the editors of IJATE, Dr. Omer Kutlu and Dr. Izzet Kara, for entrusting me with the privilege of curating this special issue. The realization of this endeavor would not have been possible without the unwavering support and meticulous efforts of the IJATE Editorial Team. Their commitment has been instrumental in bringing this special issue to fruition.

Assoc Prof. Okan Bulut
Measurement, Evaluation, and Data Science

Faculty of Education, University of Alberta
6-110 Education Centre North, 11210 87 Ave NW,
Edmonton, AB T6G 2G5 CANADA

E-mail: bulut@ualberta.ca

CONTENTS

International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 1–96
https://doi.org/10.21449/ijate.1406304
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Invited Article*

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

Bruno D. Zumbo [1*]

[1]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

*CONTACT: Bruno D. Zumbo bruno.zumbo@ubc.ca University of British Columbia, Measurement, Evaluation, & Reseach Methodology Program, Department of Educational and Counselling Psychology, and Special Education, Vancouver, BC Canada V6T 1Z4

e-ISSN: 2148-7456 /© **IJATE** 2023

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 multidisciplinarity 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 Courtis's (1921) descriptions of a test as valid if it measures what it is supposed to. Courtis 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, Courtis 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*

Hubley 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.

Hubley 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 Hubley 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 Hubley 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, Shaffner'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 Shaffner'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), Hubley 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 make 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]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" (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 $f: \Omega \to \Phi$. The image *space* Φ is thought of as the space of test-takers or survey respondents; thus, for any individual $\phi \in \Phi$, we construe $X\left(f^{-1}(\phi)\right)$ to capture all possible outcomes of the measurement process X for the particular individual ϕ. Let $\sigma(A)$ denote the usual σ-algebra generated by the generic function (or random variable) $A: \Omega \to \Lambda$; i.e.

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

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

$$\text{X} = \text{T} + \text{E}, \text{ where } \text{T} \coloneqq \mathbb{E}\left((X|\sigma(f)\right),\ f: \Omega \to \Phi. \tag{1}$$

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

$$\text{X} - \text{E} \text{ is } \sigma(f)\text{-measurable}, \text{T} \coloneqq\ \mathbb{E}\left((X|\sigma(f)\right), \mathbb{E}\left((E|\sigma(f)\right) = 0, \Omega \to \Phi. \tag{2}$$

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, Ω_I and Ω_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

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

the collection of all true scores is the Hilbert subspace

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

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 contrast with 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, Ω_I, Ω_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:

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

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

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

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; Hubley & 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, Majcen,~~

~~Schwenkmezger, and Buchner (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 as 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 explanatory-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

$$y = \Lambda \eta + \varepsilon, \tag{3}$$

where y is a (p x 1) column vector of continuous scores for person j on the p items, Λ is a (p x 1) column vector of loadings (i.e., regression coefficients) of the p items on the latent variable, η is the latent variable score for person j, and ε is (p x 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, …, C-1, define the latent variable y* such that

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

where τ_c, τ_{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 $-\infty = \tau_0 < \tau_1 < \cdots < \tau_{C-1} < \tau_C = \infty$.

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

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

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 δ 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.

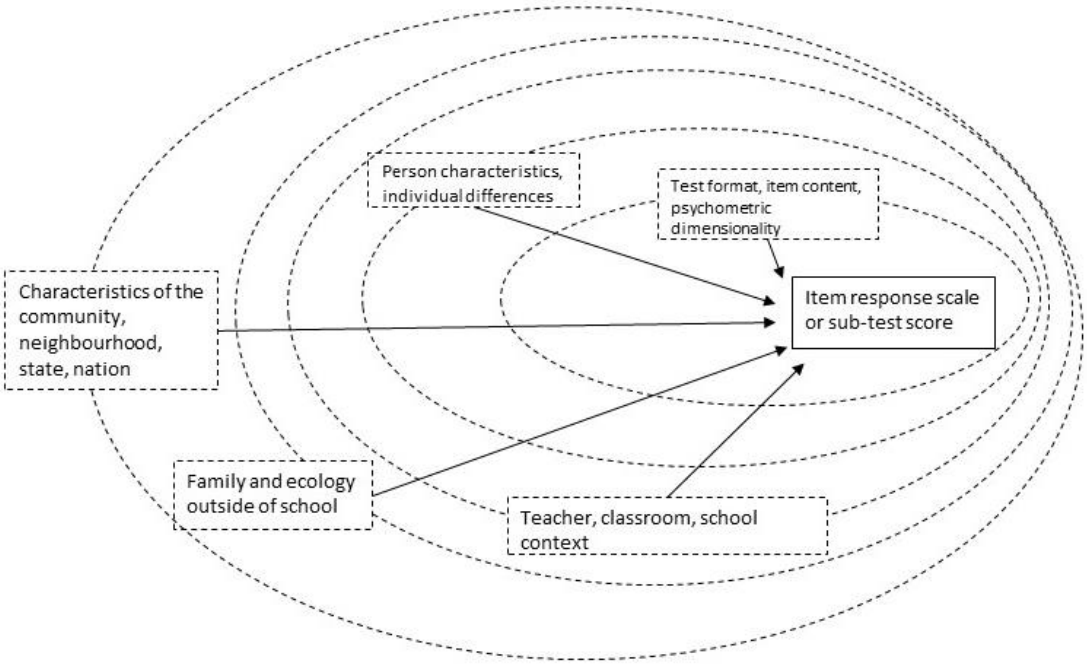
Allow me to make a critical sidebar remark before describing the ecological model. I do not wish to be interpreted as suggesting that I am the first or sole psychometrician to take aim at conventional assessment research and validation practices and offer an alternative item response model, which is 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 sub-section 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.

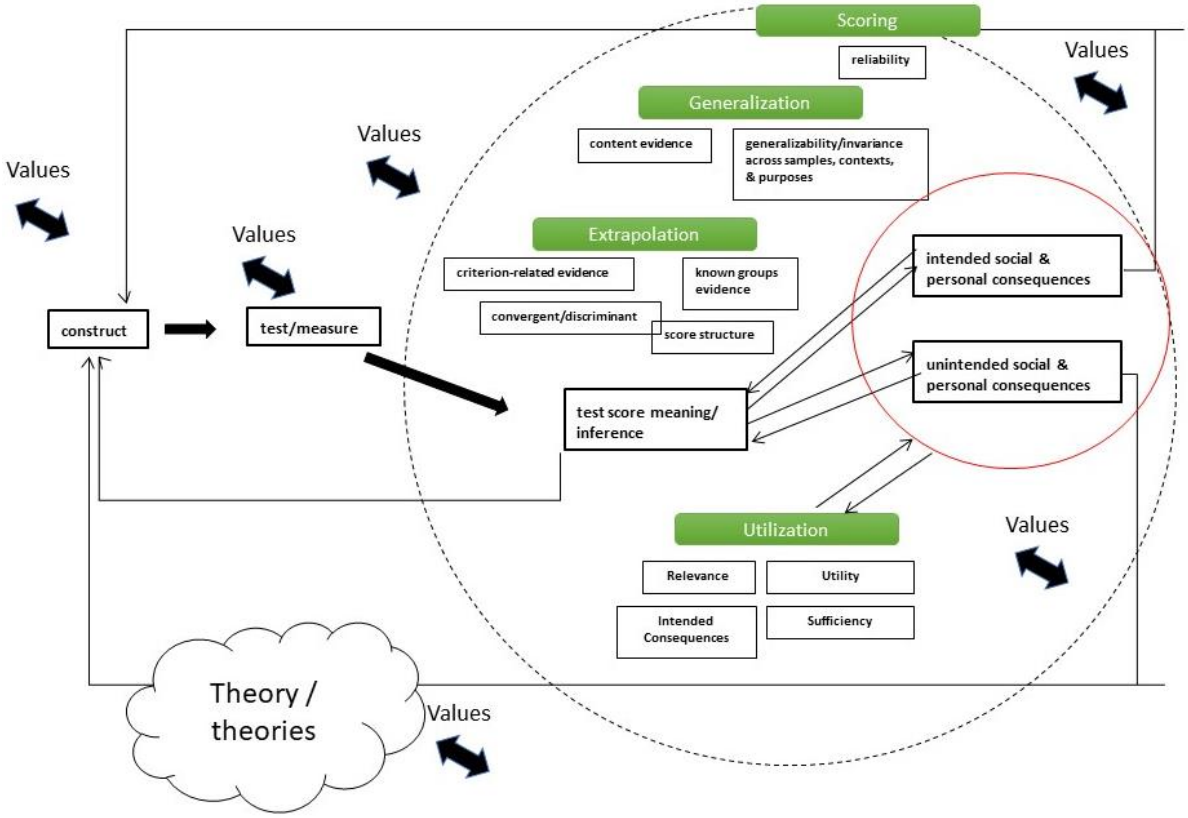
As stated by Zumbo (2023b):

> 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 Person*ality 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" (Hubley & 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* (1st 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

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., & Franić, 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 enclyclopedia 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

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, 2nd 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 Jyvaskyla Press.
Friedman, M. (1974). Explanation and scientific understanding. T*he 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* (2nd 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/10.1177/00131649921970152

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

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

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.), T*he 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 (2nd 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 (3rd 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 (2nd 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
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* (3rd 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
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 scienc*e (pp. 151–163). Springer, Cham. 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

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

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

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

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

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

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 (AEAEurope), 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

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

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

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

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.

International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 97–115
https://doi.org/10.21449/ijate.1349918
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

An investigation of factors related to collaborative problem-solving skills with mediation models

Ozge Arici[1*], **Omer Kutlu**[2]

[1]Republic of Türkiye Ministry of National Education, Ankara, Türkiye
[2]Ankara University, Faculty of Educational Sciences, Department of Educational Measurement and Evaluation, Ankara, Türkiye

ARTICLE HISTORY

Received: Aug. 25, 2023
Revised: Oct. 05, 2023
Accepted: Oct. 07, 2023

Keywords:
Mediation effect,
Multiple mediation model,
Multilevel mediation model,
PISA,
Collaborative problem solving.

Abstract: This study investigated the factors that are directly and indirectly related to collaborative problem-solving skills of students in Türkiye with multiple and multilevel mediation models, according to PISA 2015 results. The PISA 2015 Türkiye sample consisted of 5895 students. After missing data assignment and outlier analysis, the analyses were performed over the data set of 5882 students. In this study, whether the variables of valuing teamwork and valuing relationships show a mediation effect was tested with the Bootstrap method through multiple mediation models constructed with the dependent variable of collaborative problem-solving and the independent variables of school belonging and disciplinary climate. Our analyses revealed that the mediator variables had significant effects between school belonging and collaborative problem-solving. Similarly, the mediation effect between the disciplinary climate and the collaborative problem-solving was also significant. Multilevel mediation models constructed with the independent variables of students' behavior hindering learning and extracurricular creative activities were analyzed with the multilevel structural equation modeling. The findings indicated that the variables of valuing relationships and valuing teamwork did not have a significant mediation effect between extracurricular creative activities and collaborative problem-solving scores. Similarly, it was found that the mediation effect between the students' behavior hindering learning and collaborative problem-solving scores was not significant. In light of all these findings, it is recommended that school practices be strengthened to improve students' sense of belonging to school and a positive disciplinary climate, to develop students' collaborative problem-solving skills, and to improve attitudes towards collaboration.

1. INTRODUCTION

According to the Organization for Economic Co-operation and Development (OECD), cognitive skills, as well as non-cognitive skills associated with them, draw attention to education. Interpersonal and social skills which are non-cognitive skills play the role of mediators for the appearance or development of cognitive skills (Marzano & Heflebower, 2012; Kutlu & Kula-Kartal, 2018). On the other hand, school- and classroom-level features generally have a weaker effect on students' academic performance, compared to students' characteristics.

*CONTACT: Ozge ARICI oarici27@gmail.com Republic of Türkiye Ministry of National Education, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

This indicates that students' characteristics may have a direct effect on students' academic performance, while their school and classroom-level skills may have an indirect effect. However, school- and classroom-level features can have a direct effect on non-cognitive skills (like self-efficacy, motivation, etc.) and students' behavior (like skipping school, bullying, etc.) (OECD, 2017a; OECD, 2017b). In this context, the Programme for International Student Assessment (PISA) provides an important opportunity to determine the relationships between non-cognitive skills and students' academic performance.

The data obtained through PISA are used for determining the factors associated with student achievement and developing standards to increase the quality of education systems (OECD, 2017a). In PISA, each semester focuses on only one of the domains covering reading literacy, science literacy, and maths literacy. Through PISA, the OECD has made assessments in these basic areas, as well as in problem-solving and individual (creative) problem-solving in the 2003 and 2012 frameworks, and in collaborative problem-solving (CPS) in the 2015 framework. The 21st century requires acquiring high-level thinking skills, such as problem-solving, as well as an understanding of key academic content. Openness to problem-solving also affects students' academic achievement in other learning domains. Therefore, different dimensions of problem-solving skills need to be evaluated and learning environments and measurement and evaluation methods should be regulated in this direction (Kutlu et al., 2017; OECD, 2017a).

In today's world, skills such as creativity, solving complex problems, written and verbal communication, and working in collaboration have emerged as skills required for the workforce. The development of these skills requires that learning environments are organized in such a way that students are forced to communicate effectively, manage conflict, form teams, and reach a consensus on the issues necessary for living together. Schools must use the activities required by the CPS and carry out assessments and evaluations accordingly. (Kutlu & Kula-Kartal, 2018; McKenna, 2017). Making assessments for CPS skills in PISA is important in this sense.

CPS competency in PISA 2015 was defined as: "the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills, and efforts to reach that solution" (OECD, 2017a; OECD, 2017c). Individual problem-solving skills (understanding the problem content, applying problem-solving strategies, etc.) and collaboration components (cognitive and social skills to allow shared understanding, knowledge, and information flow, to create and understand an appropriate team organization, and to perform coordinated actions) have clustered within the scope of CPS (OECD, 2017a; OECD, 2017c).

Students' characteristic features such as interpersonal skills, personality traits, motives, self-efficacy perceptions, and perspectives on various issues affect their individual problem-solving and collaboration skills (Charles & Lester, 1982; Morgeson et al., 2005; Yayan, 2010). Assessments on problem-solving and individual (creative) problem-solving domains in PISA 2003 and PISA 2012 assessments provide an important source of data for determining the factors related to the problem-solving skills of students in Turkey (Aşkar & Olkun, 2005; Akyüz & Pala, 2010; Birbiri, 2014; İleritürk et al., 2017; Pala, 2008; Sertkaya, 2016). The majority of these studies that aim to determine the factors related to students' academic performance in the PISA problem-solving domain discussed the direct effects between variables. Using methods for determining the mediation effect is thought to be more effective in revealing the factors related to student achievement because of the complex relationships between variables.

Given the importance of determining CPS skills and the factors affecting these skills, it is crucial to investigate the factors related to the CPS skills of students in Turkey either directly or indirectly. The purpose of this study is to examine the factors that are directly and indirectly

related to CPS skills of students in Turkey with multiple and multilevel mediation models using PISA 2015. Within this general purpose, the following research questions are sought (the variables in the research questions are explained under the title of "Models set within the scope of the research" in the method section).

1. In the multiple mediation model, do the attitudes towards collaboration (index of valuing teamwork and index of valuing relationships) have a mediation effect on the relationship between a sense of belonging at school and CPS skills?
2. In the multiple mediation model, do the attitudes towards collaboration (index of valuing teamwork and index of valuing relationships) have a mediation effect on the relationship between disciplinary climate and CPS skills?
3. In the multilevel mediation model, do the attitudes towards collaboration (index of valuing teamwork and index of valuing relationships) have a mediation effect on the relationship between extracurricular creative activities and CPS skills?
4. In the multilevel mediation model, do the attitudes towards collaboration (index of valuing teamwork and index of valuing relationships) have a mediation effect on the relationship between students' behavior hindering learning and CPS skills?

2. METHOD

As this research aims to determine the factors related to CPS skills, the relational survey model, one of the general survey models, was used in the study (Tabachnick & Fidell, 2013).

2.1. Sample

The PISA 2015 assessment was conducted with the participation of approximately 540,000 students from 72 participating countries and economies, representing approximately 29 million students. Assessment on the CPS domain, on the other hand, was performed with the participation of approximately 125,000 students from 52 countries (OECD, 2017c). The sample of Turkey in the PISA 2015 assessment included 5895 students and 187 schools selected by cluster sampling method. In the PISA assessment, the school sample was determined by stratified random sampling method stratification according to school type and location of schools. In the second stage, on the other hand, the students to participate in the assessment in these schools were determined by random method.

2.2. Data Collection Tools

In this study, data on variables within the scope of attitudes towards collaboration such as valuing relationships and valuing teamwork variables, sense of belonging at school, and disciplinary climate were obtained from the student questionnaire under the PISA 2015 Turkey assessment, while the data on variables of students' behavior hindering learning and extracurricular creative activities were obtained from the school questionnaire. Scores related to the CPS skills were obtained from the PISA 2015 achievement test.

2.3. Models Set Within the Scope of the Research

The research aimed to determine the factors that have a direct and indirect relationship with the CPS skills of students in Turkey through multiple and multilevel mediation models. Within the scope of the study, the possible factors affecting the CPS skills of the students were determined by the literature review, taking into account the components of CPS skills, problem-solving, and collaboration components. For this purpose, multiple and multilevel mediation models were set.

PISA encompasses students' attitudes towards collaboration, which is among the non-cognitive skills thought to be related to students' achievement in the CPS domain. In PISA 2015, students' attitudes towards collaboration were examined through the index of "valuing teamwork

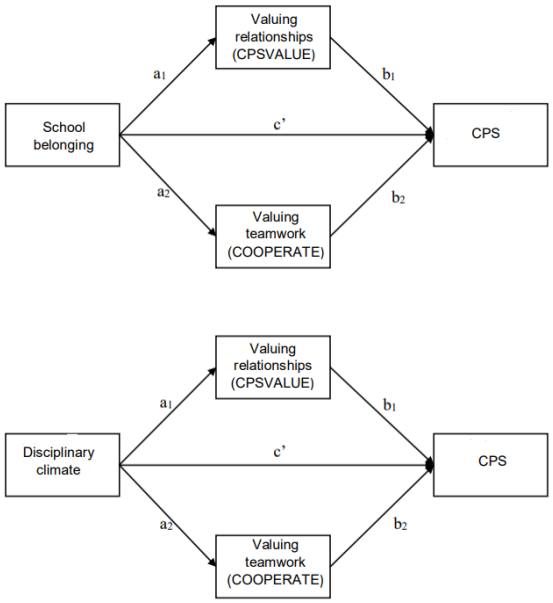
(COOPERATE)" and the index of "valuing relationships (CPSVALUE)". Valuing relationships is defined as altruistic interactions when the student engages in collaborative activities while valuing teamwork is about what teamwork can produce contrary to working alone (OECD, 2017c). The data on the attitude towards collaboration variable consists of the indexes of valuing teamwork and valuing relationships.

Students' interactions with other elements in the school are significantly related to their interpersonal skills, which are addressed within the framework of CPS skills. One of the concepts considered in this context is students' sense of belonging at school (OECD, 2017c). The sense of belonging at school gives students a sense of security, identity, and community, and all these gains could support academic, psychological, and social development (Adelabu, 2007; Anderman, 2002; Booker, 2004; Goodenow & Grady, 1993; Kutlu & Kula-Kartal, 2018; OECD, 2017d; Sarı, 2013; Sarı & Özgök, 2014).

The characteristics related to the learning environments such as teaching practices, teacher attitudes, classroom climate, competitive learning environment, classroom size, etc. were found to be associated with students' problem-solving skills (Begde, 2015; Çilingir, 2015; Ebret, 2015; Koçoğlu, 2017; Konu, 2017; Kurbal, 2015; Yayan, 2010). The PISA 2015 examined the effect of the school climate on student achievement under the heading learning environments. Schools with a good climate minimize violence, bullying, threats, and oppression, moreover, it ensures educating students who respect one another, have learned the culture of living together, and learned to be a team instead of a power struggle (Doğan, 2017). One of the variables covered by the school climate is the disciplinary climate (OECD, 2017c). A disciplined and fair learning environment helps students acquire social skills that will enable them to construct rewarding relationships at school, which they need to build with both their peers and teachers. Besides, there is a strong relationship between disciplinary climate and the sense of belonging at school (OECD, 2017e). In this regard, in the PISA 2015, students were asked about the frequency of behaviors in the classroom that hinder learning, and thus, the index of disciplinary climate was constructed.

Considering the effects of students' sense of belonging at school and their perceptions of disciplinary climate on their collaboration approaches, which are the social dimensions of CPS, mediation models were constructed. Thanks to the mediation models constructed, the effect of students' sense of belonging at school and the disciplinary climate they perceive on their CPS skills were examined through students' attitudes towards collaboration. The variable of attitude towards collaboration was addressed with the index of valuing teamwork and valuing relationships. Because there are multiple mediators, the mediation models constructed were multiple mediation models. The multiple mediation models constructed with the independent variables of sense of school belonging at school and disciplinary climate are shown in Figure 1.

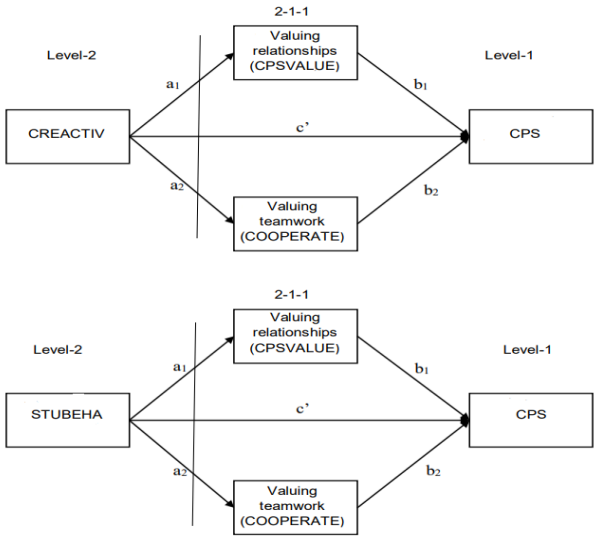
Figure 1. *Multiple mediation models constructed within the scope of research.*



The PISA 2015 assessed the effects of the school climate on both students' behavior and their academic achievements. In this context, relations between students and student activities were discussed. To determine how student behavior affects the learning environment, school principals were asked, using the school questionnaire, to what extent student truancy, skipping school, lacking respect for teachers, using alcohol or illegal drugs, and bullying other students hinder student learning. Using the data obtained in this way, the index of students' behavior hindering learning (STUBEHA) was constructed (OECD, 2017e). Activities such as sports, music and discussion groups, etc. contribute to the development of students' cognitive and non-cognitive skills. Among them, skills such as independence, compliance with guidelines, and getting on well with authority figures and peers are very important in the development of students' CPS skills in school life (and beyond) (Carneiro & Heckman, 2005; Covay & Carbonaro, 2010; Farb & Matjasko, 2012; Farkas, 2003; Howie et al., 2010, as cited in OECD, 2017c). In this sense, in the PISA 2015 school questionnaire, school principals were asked to report what extracurricular activities their schools offered to 15-year-old students, and the index of creative extracurricular activities at school (CREACTIV) was computed (OECD, 2017e).

Considering the effect of school climate on both students' behavior and academic achievement, mediation models were constructed to determine the relationship between students' CPS skills and student activities and inter-student relations. Mediators in both mediation models constructed are the variables of valuing teamwork (COOPERATE) and valuing relationship (CPSVALUE), which are the subgroups of attitude towards collaboration. Since the independent variables of extracurricular creative activities (CREACTIV) and students' behavior hindering learning (STUBEHA) collected from school principals through the school questionnaire are group-level (level-2) variables, the mediation models are multilevel mediation models. The multilevel mediation models constructed with the independent variables of extracurricular creative activities (CREACTIV) and students' behavior hindering learning (STUBEHA) are shown in Figure 2.

Figure 2. *Multilevel mediation models set within the scope of research.*



2.4. Data Analysis

Before the testing of mediation models, missing data assignment and removing outliers procedures were performed. Assignments were made for missing data using the expectation-maximization (EM) method. Following the missing data assignment and outlier analysis, the data analyses were performed over the data set of 5882 students from 187 schools.

It was checked out whether there was a multicollinearity problem, as the mediation analyses were based on regression models. In terms of the examination of the multicollinearity problem, the correlation between variables, tolerance, variance inflation factor (VIF), and condition index (CI) was also examined (Büyüköztürk, 2007; Çokluk et al., 2010; Field, 2009; Kalaycı, 2009). As a result, it is concluded that there is no multicollinearity problem.

Finally, the intra-class correlation coefficient (ICC) was examined to determine whether multilevel modeling was necessary. Tofighi and Thoemmes (2014) argue that if ICCs for level-1 variables in mediation models constructed with hierarchical data are greater than zero, then a multilevel model would give more accurate results. In the multilevel mediation models set within the scope of the research, the level-1 variables are composed of the scores of CPS, which is the dependent variable, and the indexes of valuing relationship (COOPERATE) and valuing teamwork (CPSVALUE), which are mediators. The intra-class correlation coefficient for the CPS scores was found as ρ=0.51. Accordingly, 51% of the differences between the CPS scores were due to the difference between schools, and 49% was due to the differences between students studying at the same school. Additionally, 1% of the differences regarding the students' valuing teamwork and 4% of the differences regarding their valuing relationships were due to differences between schools. In this case, it was decided to adopt a multilevel approach in the analysis of mediation models constructed with the independent variables (level-2 variables) of extracurricular creative activities and students' behavior hindering learning.

Multiple mediation models proposed for the first and second research questions were tested using the bootstrap method. The number of bootstrap samples generated within the scope of the research is 5000. While testing mediations with the bootstrap method, PROCESS macros in SPSS were used (Hayes, 2013; Preacher & Hayes, 2004).

As the data on extracurricular creative activities and students' behavior hindering learning mentioned in the third and fourth research questions were collected from school principals through the school questionnaire, these data constitute group-level data. The multilevel Structural Equation Model (MLSEM) was used to test multilevel mediation models set with

these variables. The approaches used in model estimation in MLSEM are basically categorized into two groups "within and between approach" and "full information-maximum likelihood". The MLR estimation method, which is considered under the full information-maximum likelihood approach, is used in the Mplus software (Muthen & Huberman, 2010; Heck & Thomas, 2015). MLR is robust to skewed distributions and calculates the chi-square test statistic when observations are dependent (Heck & Thomas, 2015). MLR estimation method was used in this study. The software package Mplus version 7.0 was used for MLSEM analysis.

In this study, the effect size values obtained through the ratio of indirect effect to the total effect, discussed in the section on ratio and proportion calculations, were used.

3. RESULTS

3.1. Results Regarding the Multiple Mediation Model Constructed with the Independent Variable of Sense of Belonging at School

In the multiple mediation model constructed with the independent variable of sense of belonging at school and the dependent variable of CPS scores, whether the variables of valuing teamwork and valuing relationships showed a mediation effect together was examined. In the mediation analysis, the bootstrap method was used, and the direct and total impact coefficients obtained as a result of the analysis are presented in Table 1.

Table 1. *Effect coefficients of multiple mediation model (1).*

Parameter	B	S_B	*t*	*p*
a_1	0.092	0.013	6.945	0.000*
a_2	0.050	0.011	4.507	0.000*
b_1	8.650	0.913	9.470	0.000*
b_2	2.269	1.094	2.075	0.038*
c’	5.407	0.781	6.923	0.000*
c	6.312	0.788	8.007	0.000*

*$p<0.05$

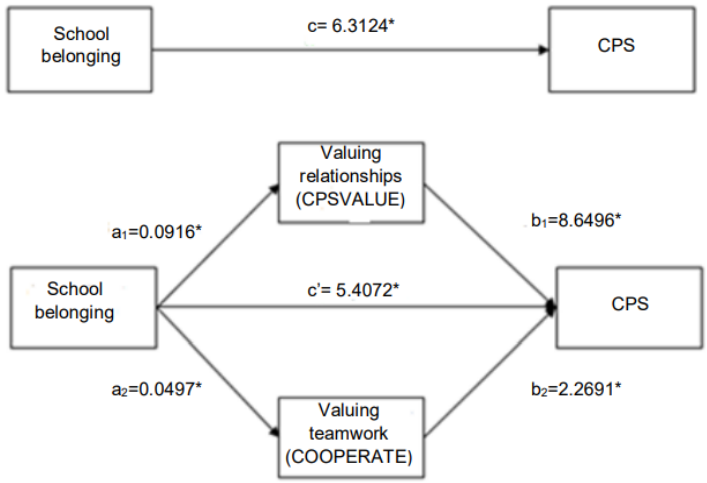
The parameters a_1 and a_2 were unstandardized regression coefficients representing the effect between the sense of belonging at school and mediators of valuing relationships and valuing teamwork, respectively. The parameters b_1 and b_2 were unstandardized regression coefficients representing the effect between the mediators of valuing relationships and valuing teamwork and CPS scores, respectively. The parameters c and c’ were parameters that indicated the total effect and direct effect between dependent and independent variables, respectively. Table 1 indicates the standard error values (SB) for the relevant effect coefficients and the t-values and significance levels for this effect. Accordingly, there was a positive and significant relationship between the sense of belonging at school and CPS scores (c=6.312, $t(5880)=8.007$, $p=0.000$). A one-unit increase in the sense of belonging at school variable caused an increase of 6.3124 in the CPS scores. Similarly, there was a positive significant relationship between the sense of belonging at school and valuing relationships ($a_1=0.092$, $t(5880)=6.945$, $p=0.000$) and valuing teamwork (a2=0.050, t(5880)=4.507, $p=0.000$). A one-unit increase in the independent variable caused an increase of 0.092 and 0.050 units in the mediators, respectively. When the effects of the mediators on the dependent variable were examined, a positive significant relationship was observed between the CPS scores and valuing relationships ($b_1=8.650$, $t(5880)=9.470$, $p=0.000$) and valuing teamwork (b2=2.269, t(5880)=2.075, $p=0.038$). A one-unit increase in the variable of valuing relationships caused an 8.6450-unit increase in the CPS scores, and the increase in the variable of valuing teamwork led to an increase of 2.269 units.

Hoyle and Kenny (1999) suggest that the power of the mediation test increases in mediation models when the coefficient b between the mediator and the dependent variable exceeds the coefficient a between the independent variable and the mediator. It is, therefore, important in

the selection of mediators, to select variables that have a relationship like b=a or b>a. As can be seen in Table 1, the effect coefficients in the first multiple mediation model constructed with the independent variable of sense of belonging at school are a_1=0.092, a2=0.050, b_1=8.650, and b2=2.269. In this case, since b>a, it can be stated that the mediators that have stronger relationships with the dependent variable compared to the independent variable are determined.

Figure 3 shows the model for the mediation effect of the variables of valuing teamwork and valuing relationships between the variables of CPS and the sense of belonging at school, and the effect coefficients in this model.

Figure 3. *Multiple mediation model (1) for the variable of attitude towards collaboration.*



When the coefficient c (6.312) representing the total effect between the sense of belonging at school variable and the CPS variable and the coefficient c' (5.407) representing the direct effect between these two variables were compared, it was seen that under the influence of mediators, the predictive power of the sense of belonging at school variable on CPS scores decreased. This reduction in the mediator effect indicated partial mediation. The fact that the independent variable is no longer a significant predictor of the dependent variable under the control of the mediator is interpreted as full mediation, whereas the fact that the independent variable is still a significant predictor of the dependent variable but the effect decreases is interpreted as partial mediation (Baron & Kenny, 1986). Zhao et al. (2010) emphasized that it is insufficient to know the statistical significance of c and c' coefficients in order to determine whether there is or not, and that a comparison should be made between the coefficients. For this reason, the c'/c ratio was analyzed. It was observed that the relevant ratio was approximately 0.86. In other words, approximately 86% of the total effect was explained by the direct effect of the variables.

After direct and total effects, the indirect effects between the variables were examined. 95% confidence intervals for indirect effects were examined with 5000 bootstrap resamples (Preacher & Hayes, 2008). Table 2 shows the relevant results.

Table 2. *Indirect effect coefficients of multiple mediation model (1).*

			95% Confidence Interval	
Parameter	Effect	SE_{Effect}	Lower	Upper
∑ab	0.905	0.167	0.590	1.258
a_1b_1	0.793	0.158	0.515	1.149
a_2b_2	0.113	0.064	0.011	0.267

Note. Bootstrap resample=5000

When Table 2 is examined, in terms of the sense of belonging at school and CPS score, it can be seen that the 95% confidence interval of the a_1b_1 indirect effect regarding the mediator of valuing relationships did not contain 0 (a_1b_1=0.793; CI= [0.515, 1.149]). It was observed that, in terms of the sense of belonging at school and CPS score, the 95% confidence interval of the indirect effect a_2b_2 related to the mediator of valuing teamwork did not contain 0, too (a_2b_2=0.113; CI= [0.011, 0.267]). Examining the 95% confidence interval for the total indirect effect between the dependent and independent variable, it was similarly found that it did not contain 0 ($\sum$ab= 0.905; CI= [0.590, 1.258]). The fact that confidence intervals for indirect effects do not contain 0 indicates that the mediation effect is confirmed (Jose, 2013; MacKinnon, 2008). In other words, the variables of valuing relationships and valuing teamwork considered within the scope of attitude towards collaboration together show a significant mediator effect in the model set with the sense of belonging at school and CPS scores.

Following the significance of the mediation effect, regarding the mediation effect of valuing teamwork and valuing relationships within the attitude towards collaboration, effect size values were obtained by the ratio and proportion approach. Accordingly, the mediation effect size, which was suggested by Jose (2013) and obtained through the ratio of indirect effect to total effect (ab/c), was computed. The ab/c ratio for the variables of valuing relationships and valuing teamwork were 0.126 and 0.018, respectively. These rates indicated that 13% of the total effect of the sense of belonging at school on the CPS scores was explained by the variable of valuing relationships, while 2% by the variable of valuing teamwork. As a result, the CPS skill scores of students who further feel a sense of belonging at school increased. 15% of this increase was explained by the fact that the sense of belonging at school increased students' attitudes towards collaboration.

3.2. Results Regarding the Multiple Mediation Model Constructed with the Independent Variable of Disciplinary Climate

In the multiple mediation model constructed with the independent variable of disciplinary climate and the dependent variable of CPS scores, it was examined whether the variables of valuing teamwork and valuing relationships showed a mediation effect together. The direct and total effect coefficients obtained in the mediation analysis are shown in Table 3.

Table 3. *Effect coefficients of multiple mediation model (2).*

Parameter	B	S_B	*t*	*p*
a_1	0.169	0.016	10.680	0.000*
a_2	0.086	0.013	6.495	0.000*
b_1	8.115	0.913	8.887	0.000*
b_2	2.236	1.089	2.052	0.040*
c'	9.160	0.945	9.694	0.000*
c	10.726	0.947	11.356	0.000*

*p<0.05

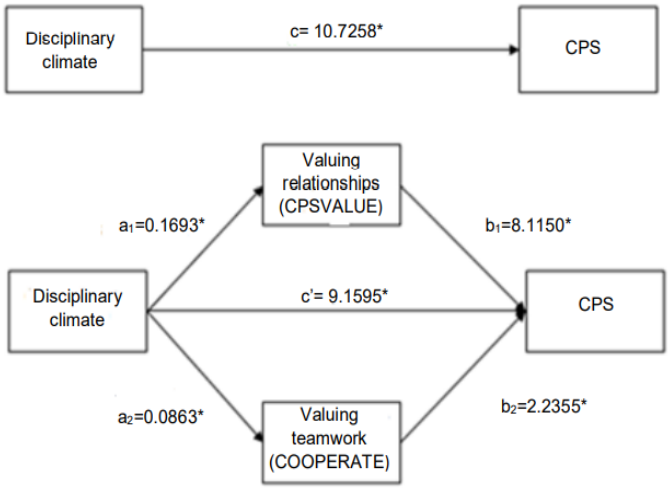
There was a positive and significant relationship between disciplinary climate and CPS scores (c=10.726, $t_{(5880)}$=11.356, p=0.000). A one-unit increase in the disciplinary climate variable caused an increase of 10.726 units in the CPS scores. Similarly, there was a positive significant relationship between disciplinary climate and valuing relationships (a_1=0.169, $t_{(5880)}$=10.680, p=0.000) and valuing teamwork (a_2=0.086, $t_{(5880)}$=6.495, p=0.000). A one-unit increase in the independent variable caused an increase of 0.169 and 0.086 units in the mediators, respectively. Examining the effects of the mediators on the dependent variable, it was observed that there was a positive significant relationship between CPS scores and valuing relationships (b_1=8.115, $t_{(5880)}$=8.887, p=0.000) and valuing teamwork (b_2=2.236, $t_{(5880)}$=2.052, p=0.040). A one-unit increase in the variable of valuing relationships caused an increase of 8.115 units in the CPS

scores, while a one-unit increase in the variable of valuing teamwork led to an increase of 2.236 units, as well.

In the multiple mediation model set with the disciplinary climate independent variable and the CPS scores dependent variable, the effect coefficients were a_1=0.169, a_2=0.086, b_1=8.115, and b_2=2.236. In this case, it can be said that, as b>a, mediators that have stronger relationships with the dependent variable compared to the independent variable are determined. This increases the power of the mediation test regarding the model established (Hoyle & Kenny, 1999).

Figure 4 shows the model for the mediation effect of the variables of valuing teamwork and valuing relationships between the variables of CPS and the disciplinary climate, and it shows the effect coefficients in this model.

Figure 4. *Multiple mediation model (2) for the variable of attitude towards collaboration.*



Comparing the coefficient c (10.726) representing the total effect between the disciplinary climate variable and the CPS variable and the coefficient c' (9.160) representing the direct effect between these two variables, it was observed that the predictive power of the disciplinary climate variable on the CPS scores decreased under the effect of the mediators. This decrease in the mediator effect points to partial mediation (Baron & Kenny, 1986). When the c'/c ratio was calculated, it was seen that 85% of the total effect was explained by the direct effect of variables. For the significance of indirect effects, 95% confidence intervals were examined with 5000 bootstrap resamples. Relevant results are given in Table 4.

Table 4. *Indirect effect coefficients of multiple mediation model (2).*

			95% Confidence Interval	
Parameter	Effect	SE_{Effect}	Lower	Upper
$\sum ab$	1.566	0.211	1.180	2.009
a_1b_1	1.374	0.211	0.995	1.841
a_2b_2	1.193	0.103	0.016	0.426

Note. Bootstrap resample=5000

When the indirect effects and confidence intervals in Table 4 were examined, it was observed that the mediation effect of the variable of valuing relationships, considered within the scope of attitude towards collaboration, between the disciplinary climate and CPS scores was confirmed. This was because the 95% confidence interval for the indirect effect a_1b_1 did not contain 0 (a_1b_1=1.374; CI= [0.995, 1.841]). It was seen that the confidence interval for the variable of valuing teamwork, considered within the scope of the attitude towards collaboration, did not contain the value 0, too (a_2b_2=1.193; CI= [0.016, 0.426]). Similarly, the confidence

interval for the total indirect effect did not contain 0 (∑ab=1.566, CI= [1.180, 2.009]). This indicated that the mediation effect of the variables of valuing relationships and valuing teamwork, considered within the scope of the attitude towards collaboration, between the disciplinary climate independent variable and the dependent variable of CPS scores was significant.

The mediation effect size values obtained through the ratio of indirect effect to total effect (ab/c) for the variables of valuing relationships and valuing teamwork were found as 0.128 and 0.018, respectively. These rates indicated that 13% of the total effect of the discipline climate on the CPS scores was explained by the variable of valuing relationships, while 2% by the indirect effect set by the variable of valuing teamwork. As a result, students' CPS scores increased in the classrooms where a more positive discipline climate dominated according to students' opinions, whereas 15% of this increase was explained by the positive effect of the positive disciplinary climate on students' attitudes towards collaboration.

3.3. Results Regarding the Multilevel Mediation Model Constructed with the Independent Variable of Extracurricular Creative Activities

In the multilevel mediation model constructed with the independent variable of extracurricular creative activities and the dependent variable of CPS scores, it was examined whether the variables of valuing teamwork and valuing relationships showed a mediation effect together. In the multilevel mediation model constructed, the level-2 variable was composed of extracurricular creative activities, while the level-1 variable was composed of valuing relationships, valuing teamwork, and collaborative problem-solving skills. The multilevel mediation model was tested with MLSEM.

Unlike other multilevel mediation analysis methods, the multilevel structural equation model provides information on model fit. However, when the studies on multilevel mediation models that used the MLSEM method were examined, it is striking that no goodness of fit index for mediation models was reported and interpreted both in methodological and applied studies (Pham, 2017; Preacher et al., 2010, 2011; Tofighi & Thoemmes, 2014). Besides, as a result of the mediation analysis made on multilevel mediation models, the Mplus 7 program did not generate any modification indices. It is erroneous to establish a direct relationship between fit indices and the accuracy of the model. Fit indices are a verification process of how far the model deviates from the data. Fit indices do not provide any evidence for the significance of the results (Millsap, 2007). When viewed from this aspect, the multilevel mediation models constructed were interpreted by their direct and indirect effect coefficients.

Table 5 shows the direct effect coefficients for the multilevel mediation model, which was constructed with the independent variable of extracurricular creative activities (CREACTIV), the dependent variable of CPS, and the mediators of valuing teamwork (CPSVALUE) and valuing relationships (COOPERATE) that were considered within the scope of attitudes towards collaboration.

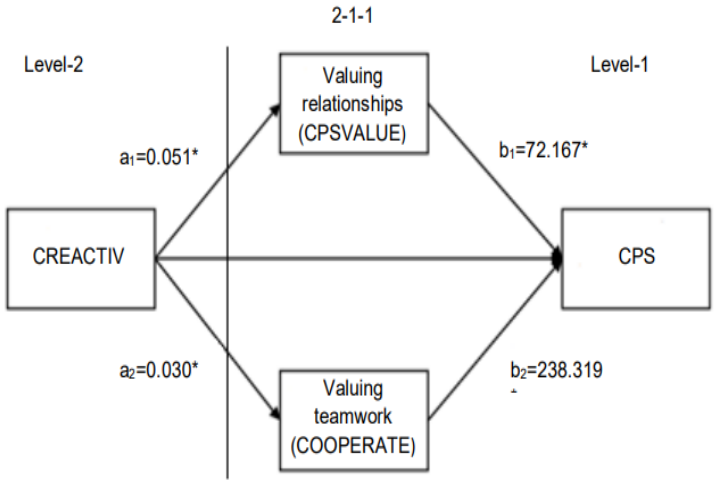
Table 5. *Direct effect coefficients for multilevel mediation model (1).*

Dependent Variable	Independent Variable	Parameter	Prediction (SE)	*p*
COOPERATE(M_1)	CREACTIV(X)	a_1	0.051 (0.022)	0.021*
CPSVALUE(M_2)	CREACTIV(X)	a_2	0.030 (0.013)	0.018*
CPS(Y)	COOPERATE(M)	b_1	72.167 (30.603)	0.018*
CPS(Y)	CPSVALUE(M_2)	b_2	238.319 (130.611)	0.068
CPS(Y)	CREACTIV(X)	c'	4.147 (6.038)	0.492

*$p<0.05$

When the relationships between the independent variable and mediators were examined in the between-group part of the model, it was seen that the variable of extracurricular creative activities (CREACTIV) predicted the variables of valuing relationships (COOPERATE) (a_1=0.051, p=0.021) and valuing teamwork (CPSVALUE) (a_2=0.030, p=0.018) significantly. A one-unit increase in the independent variable of extracurricular creative activities caused an increase of 0.051 and 0.030 units in the variables of valuing relationships and valuing teamwork, respectively. When the effect of the mediators on the dependent variable was examined, it was found that valuing relationships mediator predicted the CPS scores significantly (b_1=72.167, p=0.018), however, valuing teamwork mediator did not predict the CPS scores significantly (b_2=238.319, p=0.068). When the relationship between extracurricular creative activities under the effect of mediators and CPS scores was examined, it was found that extracurricular creative activities did not predict the CPS scores significantly (c'=4.147, p=0.492). In other words, it was observed that the direct effect of extracurricular creative activities at the between-group level on the students' CPS scores was insignificant. Figure 5 depicts the model on the mediation effect of the variables of valuing teamwork and valuing relationship between the variables of CPS and extracurricular creative activities, and it also shows the effect coefficients in this model.

Figure 5. *Multilevel mediation model (1) for the variable of attitude towards collaboration.*



The indirect effect coefficients and confidence intervals for the established mediation model are presented in Table 6.

Table 6. *Indirect effect coefficients for multilevel mediation model (1).*

				95% Confidence Interval	
Dependent Variable	Independent Variable	Parameter	Prediction (SE)	Lower	Upper
CPS(Y)	CREACTIV(X)	Indirect Effect (COOPERATE) a_1b_1	3.701 (2.283)	-0.773	8.175
CPS(Y)	CREACTIV(X)	Indirect Effect (CPSVALUE) a_2b_2	7.230 (5.153)	-2.870	17.329

The between-group indirect effects regarding the variables of valuing relationships (COOPERATE) and valuing teamwork (CPSVALUE) were 3.701 and 7.230, respectively. 95% confidence intervals for these indirect effects contained 0. This was interpreted as the variables

of valuing relationships and variable teamwork, discussed within the scope of attitude towards collaboration, did not show a significant mediation effect between extracurricular creative activities and the CPS scores.

3.4. Results Regarding the Multilevel Mediation Model Constructed with the Independent Variable of Students' Behaviour Hindering Learning

In the multilevel mediation model constructed, the level-2 variable consists of students' behavior hindering learning (STUBEHA), while level-1 variables consisted of valuing relationships (COOPERATE), valuing teamwork (CPSVALUE), and CPS scores. Table 7 shows the direct effect coefficients for the multilevel mediation model constructed.

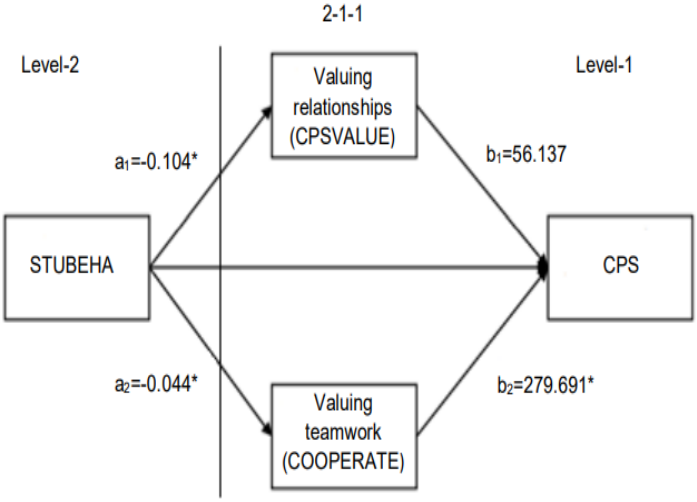
Table 7. *Direct effect coefficients for multilevel mediation model (2).*

Dependent Variable	Independent Variable	Parameter	Prediction (SE)	*p*
COOPERATE(M_1)	STUBEHA(X)	a_1	-0.104 (0.021)	0.000*
CPSVALUE(M_2)	STUBEHA (X)	a_2	-0.044 (0.016)	0.005*
CPS(Y)	COOPERATE(M)	b_1	56.137 (32.674)	0.086
CPS(Y)	CPSVALUE(M_2)	b_2	279.691 (143.706)	0.052
CPS(Y)	STUBEHA (X)	c'	-0.245 (7.166)	0.973

*p<0.05

When the relationships between the independent variable and mediators in the between-group part of the model were examined, it was observed that the independent variable of students' behavior hindering learning (STUBEHA) predicted the mediator of valuing relationships (COOPERATE) (a_1=-0.104, p=0.000) and mediator of valuing teamwork (CPSVALUE) (a_2=-0.044, p=0.005) significantly. A one-unit increase for the independent variable of students' behavior hindering learning caused a decrease of 0.104 and 0.044 units in the variables of valuing relationships and valuing teamwork, respectively. When the effect of mediators on the dependent variable was examined, it was seen that the variables of valuing relationships (b_1=56.137, p=0.086) and valuing teamwork (b_2=279.691, p=0.052) significantly did not predict the CPS scores. Examining the relationship between students' behavior hindering learning under the effect of mediators and the CPS scores, it was observed that students' behavior hindering learning did not significantly predicted the CPS scores similarly (c'=-0.245, p=0.973). Figure 6 depicts the model on the mediation effect of the variables of valuing teamwork and valuing relationship between the variables of CPS and students' behavior hindering learning, and it also shows the effect coefficients in this model.

Figure 6. *Multilevel mediation model (2) for the variable of attitude towards collaboration.*



The indirect effect coefficients and confidence intervals for the established mediation model are presented in Table 8.

Table 8. *Indirect effect coefficients for multilevel mediation model (2).*

				95% Confidence Interval	
Dependent Variable	Independent Variable	Parameter	Prediction (SE)	Lower	Upper
CPS(Y)	STUBEHA (X)	Indirect Effect (COOPERATE) a_1b_1	-5.859 (3.720)	-13.150	1.431
CPS(Y)	STUBEHA (X)	Indirect Effect (CPSVALUE) a_2b_2	-12.444 (6.986)	-26.137	1.249

The between-group indirect effects regarding the variables of valuing relationships (COOPERATE) and valuing teamwork (CPSVALUE) were -5.859 and -12.444, respectively. 95% confidence intervals for these indirect effects contained 0. This was interpreted as the variables of valuing relationships and variable teamwork, discussed within the scope of attitude towards collaboration, did not show a significant mediation effect between students' behavior hindering learning and the CPS scores.

4. DISCUSSION and CONCLUSION

In the multiple mediation model where the effect of sense of belonging at school on students' CPS skills was examined through attitudes towards collaboration, it was determined that the variable of attitude towards collaboration had a significant mediation effect between the sense of belonging at school and the CPS skills. As the students' sense of belonging at school increased, there was an increase in their CPS skills and 15% of this increase was explained by the fact that the sense of belonging at school positively affected students' attitudes towards collaboration. In the literature, there are research findings showing that there is a positive relationship between students' academic achievement and their sense of belonging at school (Adelabu, 2007; Anderman, 2002; Booker, 2004; Goodenow & Gardy, 1993; Roeser et al., 1996; Sarı, 2013; Sarı & Özgök, 2014). The influence of teachers and peers on students' sense of belonging at school is very important. Booker (2004) suggests that when students experience positive and supportive interactions with their friends and teachers, their sense of belonging at school increases. Roeser et al. (1996) found that positive interaction between teacher and student played an important role in increasing the positive effects of the school because it developed a sense of belonging at school. Students who feel accepted and approved by their peers and teachers take pleasure in attending school, in school activities and lessons more (Osterman, 2000, as cited in Sarı & Özgök, 2014). According to Adelabu (2007), students who feel a sense of belonging to school have higher levels of participation in social activities and academic work. The finding that school belonging is related to students' interactions with their teachers and peers is significantly related to interpersonal skills considered within the framework of CPS skills. Interpersonal skills are among the student characteristics that affect the achievements of individuals in collaborative problem-solving. In this context, it can be inferred that the sense of belonging at school also affects students' attitudes towards collaboration positively. Furthermore, it is thought that the increase in academic performance of students who feel a sense of belonging to the school is evidence of the effect of a sense of belonging at school on the problem-solving skills that constitute the cognitive dimension of the CPS.

The variable of attitude towards collaboration has a significant mediation effect between disciplinary climate and CPS skills. As the students' positive perceptions of the disciplinary climate in the classroom increase, there is also an increase in their CPS skills, and 15% of this increase is explained by the fact that the positive disciplinary climate positively affects students' attitudes towards collaboration. Attitudes of students in the classroom towards school and lessons, their study and listening habits, student-student and teacher-student interaction are important features that constitute the classroom climate (Erden, 1998). In this sense, the presence of a disciplined and fair learning environment in the classroom helps students acquire social skills at school that will facilitate them to establish healthy communication with their peers and teachers. In addition, there is a strong relationship between disciplinary climate and school belonging (Arum & Velez, 2012; Chiu et al., 2016; OECD, 2003, as cited in OECD, 2017e). An effective learning-teaching environment first requires individuals to communicate with each other healthily. Studies on the effect of disciplinary climate on academic achievement reveal that a positive disciplinary climate increases students' academic achievement (Akyüz & Pala, 2010; Örs-Özdil, 2017). Achieving and maintaining classroom discipline allows the teacher to spend less time on problems occurring in the classroom, concentrate more on the topics, and make lessons more effective. According to the findings obtained in this context, the fact that a positive disciplinary climate increases students' CPS scores is consistent with the situation in question. Given the effects of the disciplinary climate on the interaction between students, it can be considered that the CPS skill, in terms of its social dimension, has a positive relationship with the positive disciplinary climate. The fact that disciplinary climate affects CPS skills through attitude towards collaboration is thought to depend on students' ability to communicate with each other effectively in disciplined and fair learning environments and on not experiencing negativity in the division of responsibility for a task. Such an educational environment enables educating students who respect one another, have learned the culture of living together, and learned to be a team instead of a power struggle.

In multilevel mediation models where the effect of the independent variables of extracurricular creative activities and students' behavior hindering learning on the CPS skills are examined through the variable of attitude towards collaboration, on the other hand, it was determined that the attitude towards collaboration variable did not have a significant mediation effect. However, there is a negative relationship between students' behavior hindering learning and students' CPS skills, while there is a significant positive correlation between extracurricular creative activities and CPS skills. Students who spend more time at school through extracurricular and social activities can internalize their sense of belonging as well as improve their communication with each other. In this regard, extracurricular creative activities examined under PISA are effective in the development of students' social skills and academic achievements in schools, and thus, students' behavior hindering learning such as absenteeism, truancy, bullying, lack of respect, etc. are prevented (OECD, 2017c). Research on the characteristics of active schools indicates that a regular, supportive and positive environment in these schools. Activities such as sports, music and discussion clubs, etc. have an important role in building a supportive and positive school climate. These activities also enable students to develop their skills such as independence, compliance with guidelines, getting on well with authority figures and peers, etc. From this point of view, such activities carried out in schools are very important for the development of students' CPS skills, as they include the dimensions of leadership, communication and collaboration (OECD, 2017c). Besides, in a safe and healthy school climate, when negative behaviors such as violence, bullying, threats, and oppression are minimized, then it would be possible to educate students who respect one another, have learned the culture of living together, and manage to be a team instead of power struggle (Doğan, 2017). In this sense, it is thought that a decrease in negative student behavior would affect the students' CPS skills due to its social dimension. However, the findings obtained reveal that

extracurricular creative activities and students' behavior hindering learning, which was discussed within the scope of the research, did not have a direct and indirect significant effect on students' CPS skills through attitude towards collaboration. The probable reasons for this outcome may be due to the psychological characteristics of the variables in the mediation models constructed or the limitation of the number of mediators. Another probable reason may be the limitations of the PISA 2015 Turkey sample. In addition, the literature is open for improvement in terms of both methodological and application-oriented research for the use of MLSEM in mediation analysis.

Declaration of Conflicting Interests and Ethics

The authors declare 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 authors.

Authorship Contribution Statement

The authors contributed equally to all the stages of the study.

Orcid

Ozge ARICI https://orcid.org/0000-0003-0182-6739
Omer KUTLU https://orcid.org/0000-0003-4364-5629

REFERENCES

Adelabu, D.D. (2007). Time perspective and school membership as correlates to academic achievement among African American adolescents. *Adolescence, 42*(167), 525-538.

Akyüz, G., & Pala, N.M. (2010). The effect of student and class characteristics on mathematics literacy and problem solving in PISA 2003. *İlköğretim Online, 9*(2), 668-678. https://www.ilkogretim-online.org/fulltext/218-1596890524.pdf?1619615302

Anderman, E.M. (2002). School effects on psychological outcomes during adolescence. *Journal of Educational Psychology,94*(4), 795-809. https://doi.org/10.1037/0022-0663.94.4.795

Aşkar, P., & Olkun, S. (2005). PISA 2003 sonuçları açısından okullarda bilgi ve iletişim teknolojileri kullanımı [Use of information and communication technologies in schools in terms of PISA 2003 results]. *Eurasian Journal of Educational Research,* 19, 15-34.

Baron, R.M., & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Begde, Z. (2015). Öğretmen ve ebeveyn tutumlarının okul öncesi dönem çocuklarının problem çözme becerilerine etkisinin incelenmesi *[Investigating the effect of teacher and parent attitudes on preschool children's problem solving skills]* [Unpublished master's thesis]. Karabük Üniversitesi, Karabük, Türkiye.

Birbiri, D. (2014). *PISA 2003 ve PISA 2012 sınav sonuçlarının problem çözme becerilerine yönelik değişkenlerinin Türkiye açısından incelenmesi [Examination of variables related to problem solving skills in PISA 2003 and PISA 2012 results in terms of Turkey]* [Unpublished master's thesis]. Atatürk Üniversitesi, Erzurum, Türkiye.

Booker, K.C. (2004). Exploring school belonging and academic achievement in African American adolescent. *Curriculum and Teaching Dialogue, 6*(2), 131-143.

Büyüköztürk, Ş. (2007). *Sosyal bilimler için veri analizi el kitabı* (7. Baskı) *[Handbook of data analysis for social sciences (7th ed.]*. Pegem Akademi.

Charles, R., & Lester, F. (1982). *Teaching problem solving: What, why & how*. Dale Seymour Publicatios.

Çilingir, E. (2015). *Gerçekçi matematik eğitimi yaklaşımının ilkokul öğrencilerinin görsel matematik okuryazarlığı düzeyine ve problem çözme becerilerine etkisi [The effect of realistic mathematics education approach on primary school students' visual mathematics literacy level and problem solving skills]* [Unpublished master's thesis]. Çukurova Üniversitesi, Adana, Türkiye.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LİSREL uygulamaları.* Pegem Akademi.

Doğan, S. (2017). Okul yönetimi [School management]. Celal Tayyar Uğurlu (Ed.). *Okul kültürü ve iklimi* (s. 92-119) *[School culture and school climate]*. Anı Yayıncılık.

Ebret, A. (2015). *Etkinlik temelli matematik öğretiminin 3. sınıf öğrencilerinin problem çözme becerilerine ve matematiğe ilişkin tutumlarına etkisi [The effect of activity-based mathematics teaching on 3rd grade students' problem solving skills and attitudes towards mathematics]* [Unpublished master's thesis]. Necmettin Erbakan Üniversitesi, Konya, Türkiye.

Erden, M. (1998). *Öğretmenlik mesleğine giriş* [*Introduction to teaching profession*]. Alkım Yayınları.

Field, A. (2009). *Discovering statistics using SPSS* (3rd Edition). Sage Publication.

Goodenow, C., & Grady, K.E. (1993). The relationship of school belonging and friends' values to academic motivation among urban adolescent students. *The Journal of Experimental Education, 62* (1), 60-71. https://doi.org/10.1080/00220973.1993.9943831

Hayes, A.F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* Guilford Press.

Heck, R.H., & Thomas, S.L. (2015). *An introduction to multilevel modeling techniques* (Third edition). Routledge.

Hoyle, R.H., & Kenny, D.A. (1999). Sample size, reliability and tests of statistical mediation. In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 195-222). Sage.

İleritürk, D., Ercoşkun, N.Ç., & Kıncal, R.Y. (2017). Farklı ülkelerin PISA 2012 problem çözme becerileri sonuçlarının karşılaştırılması [Comparison of PISA 2012 problem solving skills results of different countries]. *Akademik Sosyal Araştırmalar Dergisi, 5*(43), 406-422. http://dx.doi.org/10.16992/ASOS.12023

Jose, P.E. (2013). *Doing statistical mediation & moderation*. Guilford Publicatios.

Kalaycı, Ş. (2009). *SPSS uygulamalı çok değişkenli istatistik teknikleri* (4. Baskı) [*Multivariate statistical techniques with SPSS* (4th ed.)]. Asil Yayın Dağıtım.

Koçoğlu, A. (2017). *Fen bilimleri ve matematik öğretmenlerinin özerklik desteğinin ortaokul öğrencilerinin eleştirel düşünme eğilimi ve problem çözme becerileri algısına katkısının incelenmesi [Examining the contribution of science and mathematics teachers' autonomy support to middle school students' critical thinking disposition and problem-solving skills perception]* [Unpublished master's thesis]. Mersin Üniversitesi, Mersin, Türkiye.

Konu, M. (2017). *Yaşam temelli probleme dayalı öğretim uygulamalarının öğrencilerin biyoloji dersindeki başarılarına, tutumlarına, motivasyonlarına ve problem çözme becerilerine etkisi etkisi [The effect of life-based problem-based teaching practices on students' achievement, attitudes, motivation and problem-solving skills in biology course]* [Unpublished doctoral thesis]. Atatürk Üniversitesi, Erzurum, Türkiye.

Kurbal, M.S. (2015). *6. sınıf zekâ oyunları dersi öğrencilerinin problem çözme stratejilerinin ve akıl yürütme becerilerinin incelenmesi* [Unpublished master's thesis]. Middle East Technical University, Ankara, Türkiye.

Kutlu, Ö., Kula-Kartal, S., & Şimşek, T. (2017). Identifying the relationships between perseverance, openness to problem solving, and academic success in PISA 2012 Turkey.

Journal of Educational Sciences Research, 7(1), 263-274. https://dergipark.org.tr/tr/download/article-file/698144

Kutlu, Ö., & Kula-Kartal, S. (2018). The prominent student competences of the 21st century education and the transformation of classroom assessment. *International Journal of Progressive Education, 14*(6), 69-82. https://doi.org/10.29329/ijpe.2018.179.6

Marzano, R.J., & Heflebower, T. (2012). *Teaching and assessing 21st century skills*. Marzano Research.

McKenna, J. (Sep 14, 2017). *Collaborative problem solving in the classroom.* https://robomatter.com/blog-collaborative-problem-solving/

Millsap, R.E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences, 42*(5), 875-881. https://doi.org/10.1016/j.paid.2006.09.021

Morgeson, F.P., Reider, M.H., & Campion, M.A. (2005). Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge. *Personnel Psychology, 58*(3), 583-611. https://doi.org/10.1111/j.1744-6570.2005.655.x

Muthén, L.K., & Muthén, B.O. (1998-2010). *Mplus user's guide* (Sixth Edition). Muthén & Muthén.

OECD. (2017a). *PISA 2015 assessment and analytical framework: science, reading, mathematic, financial literacy and collaborative problem solving* (Revised Edition). PISA, OECD Publishing.

OECD. (2017b). *PISA 2015 technical report.* PISA, OECD Publishing.

OECD. (2017c). *PISA 2015 results (Volume V): Collaborative problem solving*. PISA, OECD Publishing.

OECD. (2017d). *PISA 2015 results (Volume III): Students' well-being.* PISA, OECD Publishing.

OECD. (2017e). *Policies and practises for succesful schools (Volume II).* PISA, OECD Publishing.

Örs-Özdil, S. (2017). *Tekli ve çoklu aracılık modellerinde aracı değişken etkisinin bk, sobel, bootstrap yöntemleriyle karşılaştırılması (PISA 2012 matematik okuryazarlığı*) [*Comparison of mediator variable effect in single and multiple mediation models with bk, sobel, bootstrap methods (PISA 2012 mathematics literacy)*] [Unpublished doctoral thesis]. Ankara Üniversitesi, Tez No. 468272. https://dspace.ankara.edu.tr/xmlui/bitstream/handle/20.500.12575/73424/468272.pdf?sequence=1&isAllowed=y

Pala, N.M. (2008). *PISA 2003 sonuçlarına göre öğrenci ve sınıf özelliklerinin matematik okuryazarlığına ve problem çözmeye etkisi* [Unpublished Master Thesis, Balıkesir Üniversitesi]. https://hdl.handle.net/20.500.12462/1679

Pham, T.V. (2017). *The performance of Multilevel Structural Equation Modeling (MSEM) in comparison to Multilevel Modeling (MLM) in multilevel mediation analysis with non-normal data* [Doctoral dissertation, University of South Florida]. University of South Florida Graduate Theses and Dissertations. http://scholarcommons.usf.edu/etd/7077

Preacher, K.J., & Hayes, A.F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments & Computers, 36*(4), 717-731.

Preacher, K.J., & Hayes, A.F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879-891. https://doi.org/10.3758/BRM.40.3.879

Preacher, K.J., Zyphur, M.J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*(3), 209-233. https://doi.org/10.1037/a0020141

Preacher, K.J., Zyphur, M.J., & Zhang, Z. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling, 18*(2),161-182. https://doi.org/10.1080/10705511.2011.557329

Roeser, R.W., Midgley, C., & Urdan, T.C. (1996). Perceptions of the school psychological environment and early adolescents' psychological and behavioral functioning in school: The mediating role of goals and belonging. *Journal of Educational Psychology, 88*(3), 408-422. https://doi.org/10.1037/0022-0663.88.3.408

Sarı, M., & Özgök, A. (2014). Ortaokul öğrencilerinde okula aidiyet duygusu ve empatik sınıf atmosferi algısı. *Gaziantep University Journal of Social Sciences, 13*(2), 479-492. https://doi.org/10.21547/jss.256833

Sarı, M. (2013). Sense of school belonging among high school students. *Anadolu University Journal of Social Sciences, 13*(1), 147-160.

Sertkaya, V. (2016). *The relationship between student and teacher related factors and students' problem solving skill throughout Turkey and across school types: PISA 2012 analysis* (Publication No. 10128159) [Master's Thesis, Bilkent University]. Bilkent University Institutional Repository. http://hdl.handle.net/11693/32473

Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics* (5th Edition). Allyn & Bacon/Pearson Education.

Tofighi, D., & Thoemmes, F. (2014). Single-level and multilevel mediation analysis. *Journal of Early Adolescence, 34*(1), 93-119. https://doi.org/10.1177/0272431613511331

Yayan, B. (2010). *Altıncı sınıf Türk öğrencilerinin problem çözme becerilerini etkileyen öğrenci ve öğretmen özellikleri [Student and teacher characteristics affecting sixth grade Turkish students' problem-solving skills]* [Unpublished doctoral thesis]. Middle East Technical University, Ankara, Türkiye.

Zhao, X., Lynch, J.G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, 37, 197- 210.

International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 116–131
https://doi.org/10.21449/ijate.1321061
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

A data pipeline for e-large-scale assessments: Better automation, quality assurance, and efficiency

Ryan Schwarz[1], **H. Cigdem Bulut**[2*], **Charles Anifowose**[3]

[1]Vretta Inc, Toronto, ON Canada
[2]Northern Alberta Institute of Technology, Education Insights, Data & Research, Edmonton, AB Canada
[3]Vretta Inc, Toronto, ON Canada

ARTICLE HISTORY

Received: June 30, 2023
Revised: Oct. 29, 2023
Accepted: Nov. 20, 2023

Keywords:
Data pipelines,
Psychometric analysis,
Large-scale assessments,
Data validation,
Reporting.

Abstract: The increasing volume of large-scale assessment data poses a challenge for testing organizations to manage data and conduct psychometric analysis efficiently. Traditional psychometric software presents barriers, such as a lack of functionality for managing data and conducting various standard psychometric analyses efficiently. These challenges have resulted in high costs to achieve the desired research and analysis outcomes. To address these challenges, we have designed and implemented a modernized data pipeline that allows psychometricians and statisticians to efficiently manage the data, conduct psychometric analysis, generate technical reports, and perform quality assurance to validate the required outputs. This modernized pipeline has proven to scale with large databases, decrease human error by reducing manual processes, efficiently make complex workloads repeatable, ensure high quality of the outputs, and reduce overall costs of psychometric analysis of large-scale assessment data. This paper aims to provide information to support the modernization of the current psychometric analysis practices. We shared details on the workflow design and functionalities of our modernized data pipeline, which provide a universal interface to large-scale assessments. The methods for developing non-technical and user-friendly interfaces will also be discussed.

1. INTRODUCTION

The field of education is significantly influenced by the impact of testing, as evidenced by the widespread adoption of national and provincial assessment levels by various countries, alongside their active participation in international large-scale assessments. National assessment programs mostly aim to understand how well students perform in terms of curriculum expectations and standards, as well as to promote performance accountability (Volante & Ben Jaafar, 2008). On the other hand, international assessments allow countries to compare across education systems or to identify their relative strengths and weaknesses based on student performance (Addey & Sellar, 2018). Despite their distinct purposes, both national and international assessments have emerged as crucial tools for enhancing educational systems (Kamens & McNeely, 2010).

*CONTACT: H. Cigdem Bulut haticeb@nait.ca Northern Alberta Institute of Technology, Education Insights, Data & Research, Edmonton, AB Canada

e-ISSN: 2148-7456 /© **IJATE** 2023

Both national and international measurement practices have changed significantly over the past two decades. Pushes towards modernization have been supported by recent advances in both online and offline technologies applicable to the education industry. Accordingly, assessment design, delivery, scoring, and reporting methods have evolved significantly (Zenisky & Sireci, 2002). Since the first decade of the twenty-first century, numerous large-scale tests have switched from paper to computer-based administration (i.e., online tests and online assessments), becoming the standard in modernized educational programs (Moncaleano & Russell, 2018). Online assessments have been adopted more rapidly due to increased access to information and communication technologies in classrooms, technological advancements in testing, and methodological improvements in psychometrics that enable efficient, personalized assessments (Moncaleano & Russell, 2018). Moreover, the recent safety and health concerns brought on by the pandemic (Lynch, 2022) have further prompted educational institutions to embrace online assessments, ensuring both test security and the well-being of students.

Online assessments have provided great advantages such as increasing test efficiency, enabling faster and more efficient scoring and reporting, as well as improving the standardization of assessments, and enhancing test security (Wise, 2018). The modernization of assessments has improved the efficiency of scoring not only for selected-response items but also for open-response items (Liu et al., 2014; Sung et al., 2017). Utilizing these advancements in assessments has led to a decrease in time, labour, and financial costs in scoring item responses (Moncaleano & Russell, 2018).

The adoption of computer-based administration of assessments has also led to the development of various new item types referred to as technology-enhanced items (TEI) (Scalise & Gifford, 2006; Bryant, 2019). These items allow educational practitioners to enhance the extent to which test tasks reflect the knowledge, skills, and abilities of interest and to be more flexible (Scalise & Gifford, 2006; Russell, 2019). These are especially useful as it can be difficult to measure complex and high-level capabilities with traditional paper-and-pencil assessments (Zenisky & Sireci, 2002).

Inevitably, online assessments and overall modernization have brought a number of challenges to educational organizations and testing companies. The complex designs of the assessments, scoring various types of items, and ensuring the validity, reliability, and security of the assessment results necessitate meticulous planning and execution in each step of the administration. The increasing volume of large-scale assessment data also challenges organizations to effectively manage, score, and analyze data (Rutkowski et al., 2010). The difficulties begin with data storage and extend all the way to sharing/transferring results. Furthermore, feeding the sheer size of large-scale assessment data for analysis makes it difficult to proceed timely and efficiently.

To address these challenges, we have designed and implemented a modernized data pipeline that allows psychometricians and statisticians to efficiently manage the data, conduct psychometric analysis, generate technical reports, and perform quality assurance to validate the required outputs. A data pipeline itself is a series of data processing steps that begins with extracting raw data sets, processing the information, and managing that data in a systematic way, and then generating outputs at the end (Skiena, 2017). In education, data pipelines are utilized in order to develop early warning systems, predict student performance, and in data modeling for educational stakeholders (Ansari et al., 2017: Bertolini et al., 2021; Bertolini et al., 2022; Schleiss et al., 2022). As of the time of this paper, to the best of the authors' knowledge, no publicly reported project has focused on the development of a comprehensive psychometric data pipeline for large-scale educational assessments. Our work seeks to address this gap by presenting a meticulously designed and well-documented pipeline solution that caters to the specific needs of this critical domain.

The data pipeline proposed in this paper offers a fully automated, end-to-end, configurable, and customizable application, delivering psychometric analysis and data quality verification to stakeholders. It provides the preparation of assessment data for psychometric analysis based on classical test theory (CTT) and item response theory (IRT), producing CTT and IRT reports. It has proven to scale with large databases, decrease human error by reducing manual processes, efficiently make complex workloads repeatable, ensure high quality of the outputs, and reduce overall costs of psychometric analysis of large-scale assessment data. The customizable and dynamic nature of the pipeline enables the standard analysis workflow to take place in a significantly reduced time as compared to traditional practices. Verification reports are also generated, providing quality assurance and flagging errors or warnings that are brought to the immediate attention of psychometricians and statisticians. Lastly, the pipeline empowers stakeholders by offering them an interface to independently execute the entire administration process. This interface enables stakeholders to navigate through the necessary steps and perform various tasks within the pipeline without requiring extensive technical expertise. By providing this capability, stakeholders gain greater control and autonomy over the administration process, facilitating efficient and independent management of reporting requirements.

In summary, our approach offers a valuable solution for researchers and practitioners seeking versatility, reproducibility, and rigorous documentation large-scale assessment data needs. It addresses a crucial need in operational settings where manual, fragmented processes are prevalent. Our data pipeline efficiently handles diverse large-scale assessments, producing detailed analyses, psychometric reports, verification reports, and scorecards within 40-50 minutes, streamlining the entire workflow.

1.1. Psychometric Analysis

Psychometric analysis can be considered one of the most technical aspects of assessments as it requires expertise and training in educational statistics and measurement, intensive and collaborative work with subject matter experts, and the ability to comprehend and reflect educational policies in assessments. The primary measurement frameworks for psychometric analysis are CTT and IRT (Lord & Novick, 1968; Embretson & Reise, 2000). These two frameworks differ significantly in terms of complexity, assumptions, and measurement precision (Hambleton et al., 1991). In CTT, all items make an equal contribution to student scores, and item and test-taker statistics are sample-dependent (Embretson & Reise, 2000; Reise et al., 2005). By contrast, IRT analysis estimates the probability of answering an item correctly by considering student latent abilities and item parameters (Hambleton et al., 1991; Embretson & Reise, 2000; Reise et al., 2005). Therefore, the resulting item and person statistics are sample-independent, especially in non-Rasch models (Hambleton et al., 1991; Embretson & Reise, 2000; Reise et al., 2005). An IRT model estimates abilities by utilizing the pattern of item responses, whereas CTT ignores these patterns. Therefore, measurement precision becomes higher in IRT models (Hambleton et al., 1991; Embretson & Reise, 2000; Zenisky & Sireci, 2002). Although CTT provides important information to evaluate and improve the items and tests, it falls short of meeting the needs of modernized assessments in many aspects (see for further discussion, Embretson & Reise, 2000).

With the modernization of assessments, methodological changes were made in the design of the assessments and item scoring. As larger and more detailed datasets allow for more complex psychometric analysis, IRT-based analysis has been commonly used in large-scale assessments and fulfills the criteria of large-scale assessments in terms of validity and fairness (Oranje & Kolstad, 2019; Camara & Harris, 2020). As tailoring administered items to each individual produces greater measurement precision (Hambleton et al., 1991; Embretson & Reise, 2000; Zenisky & Sireci, 2002), IRT-based assessments can yield more robust results owing to the

invariance assumptions inherent in IRT, as compared to assessments based on CTT. These invariance assumptions allow for a more precise understanding of the latent traits being measured. As item and person statistics are on the same scale in IRT models, IRT provides more flexibility to testing organizations in many steps, such as adaptive testing, form building, and the expansion and maintenance of item pools (Hambleton et al., 1991). Furthermore, considering test fairness and security, educational organizations and testing companies tend to generate IRT-based test forms (Oranje & Kolstad, 2019).

1.2. Psychometric Software and Programming Languages

As psychometric methodology increases in complexity, software programs must evolve to meet the changing criteria and demands stemming from educational policies, curriculum, and testing specifications. Many new tools have been built to better design assessments, as well as understand and analyze assessment data. Table 1 shows the most commonly used psychometric software and programming languages in testing companies and educational institutions.

Table 1. *Most common psychometric software and programming languages used.*

Software	Functionality	Open-source
BILOG, MULTILOG PARSCALE	IRT applications (calibration, equating, linking)	No
WINSTEPS, BIGSTEPS	Item calibration based on Rasch Measurement and Rasch Analysis	No
IRTPRO, flexMIRT	Item calibration using IRT	No
SAS	Item calibration and test scoring using IRT	No
Mplus	Item calibration using IRT, Structural Equation Modelling	No
R	IRT applications (calibration, equating, linking, form building, CAT applications) MIRT (and unidimensional mirt), GRM, CDM, SEM, SEM, DIF, EDM, Confirmatory and Exploratory Factor Analysis Automated Test Assembly	Yes
Python	Item calibration using IRT, MIRT, GRM, CAT, CDM, SEM, G-DINA	Yes
Julia	Structural Equation Modelling Automated Test Assembly Item calibration	Yes

As shown in Table 1, it is possible to conduct various psychometric analyses with different software or programming languages. However, not all of them are able to perform analyses based on different measurement frameworks, including CTT, IRT, generalizability theory, and Rasch measurement theories. Nor can they conduct every application of IRT, such as calibration, equating, multigroup analysis, and explanatory modeling.

The primary reason that R is currently at an advantage is due to its orientation towards data and statistical analysis (Desjardins & Bulut, 2018). Psychometric and statistical-oriented packages

are typically built off of academic research and provide a reference to associated documentation in the CRAN (Comprehensive R Archive Network) library or a peer-reviewed paper. Furthermore, R provides the most versatility in terms of measurement frameworks and IRT applications thanks to the numerous packages available (Schumacker, 2019). The R programming language has also grown in popularity in the field of educational measurement (Desjardins & Bulut, 2018). One possible reason for this is that R is free/libre software and therefore incurs no costs for its use (R Core Team, 2022). The trade-off with free and open-source software is the loss of technical support from purchasing licensed applications but gaining a great amount of customizability. Additionally, these applications are fairly rigid in how they require data to be input, whereas R can be customized at the ground level to data models.

Some software packages such as BILOG, MULTILOG, and PARSCALE (du Toit, 2003: Muraki & Bock, 2003; Thissen et al., 2003) necessitate a specific format for the input data for which users need to follow a guideline (Croudace et al., 2005). As a result, traditional psychometric software presents barriers, such as a lack of functionality for managing data. When dealing with large amounts of assessment data, it is possible to run into memory issues even in commonly used data management software such as Excel with a maximum limit of 4GB (Microsoft Corporation, 2018) or IBM SPSS Statistics (IBM, 2020). Secondly, none of these software packages provide the ability to perform all item- and test-level analyses required by modernized large-scale assessments (Rupp, 2003). This means that the data preparation process often requires the use of distinct software programs, each serving a specific purpose. This necessitates the creation of input data sets tailored to individual software requirements, as well as the careful formatting and customization of outputs to meet specific needs. These tasks demand meticulous attention to detail and consume valuable time. Moreover, the repetitive nature of these steps, coupled with the manual integration of various software programs, can result in time-consuming and inefficient workflows, hampering the completion of comprehensive analyses.

Standard assessment practices involve repeating psychometric processes numerous times until adequate results are achieved. However, the repetitive nature of these manual steps poses challenges for educational practitioners, particularly psychometricians, as it increases the risk of errors. Because assessments typically have tight deadlines, completing the psychometric work on time while allowing for quality control is essential to ensure that the results are technically accurate and reliable.

Another crucial aspect to consider is the cost associated with the use of property software, which can be quite expensive (Martinková & Drabinová, 2018). This becomes particularly significant when considering the need for multiple licenses to facilitate a comprehensive analysis. Additionally, there are various challenges involved in accommodating diverse assessment requirements, such as managing exceptions and addressing unforeseen data errors. Consequently, these challenges can contribute to substantial costs in order to attain the desired research and analysis outcomes.

1.3. Reporting

Reporting is another important aspect of large-scale assessments (Ysseldyke & Nelson, 2002). Once the analyses are complete, they should be reported and shared with different stakeholders. Reports may include raw scores, proficiency levels, percentiles, and standard scores, whereas reports related to items and tests provide statistics and information at the item and test level (e.g., Goodman & Hambleton, 2004). The main aims of these reports are to deliver student outcomes and evaluate the performance items and tests. Furthermore, these reports can be utilized in order to share information with students, teachers, families, and educational policymakers (Rutkowski et al., 2010).

Reports should include clear statements for the intended educational stakeholders (Ysseldyke & Nelson, 2002). Therefore, educators may be burdened by unclear and disorganized results. As traditional software programs print standard results in a text format or proprietary formats (du Toit, 2003: Muraki & Bock, 2003; Thissen et al., 2003), manually and separately preparing these reports would entail a laborious and time-intensive endeavor. Therefore, producing customized reports would be helpful in working efficiently with many internal and external stakeholders.

2. METHOD

2.1. Design Philosophy

The underlying design philosophy of the pipeline primarily adopts a stage-oriented approach, differing from a modular approach where each major functionality of the pipeline is treated as a distinct and independent element capable of operating autonomously. The sequential nature of data processing requirements lent itself to this method, as the pipeline would need to perform various tasks before analysis and reporting. We implemented a strategy of isolating stages to ensure the separation of processing rules and the preservation of original data for comparison and validation purposes.

Although this approach is sequential in nature, the philosophy behind the pipeline was still to be both dynamic and automated. Each function was designed to handle data from any assessment with any configurations. The code therefore incorporates adaptability and atomism, eliminating the need for code replication.

2.2. Tools Used

The language chosen at the outset of the project was R (R Core Team, 2022). The R language has a few advantages over other languages considered, including Python (Van Rossum & Drake, 1995) and Julia (Bezanson et al., 2012). The main rationale behind this is that R is developed by statisticians, and as a result, its user community predominantly consists of professionals and academics from relevant fields. Consequently, there exists a high level of support for psychometric tasks within the community. Furthermore, R offers robust reporting tools, such as RMarkdown (Allaire et al., 2022), which greatly enhance the language's capabilities in generating comprehensive reports. These advantages meant the project could more efficiently get up and running by using already built open-source packages. In this pipeline, we leverage several key R packages to enhance our data analysis and reporting capabilities. Packages used in the pipeline are shown in Table 2.

Table 2. *Commonly used packages in the pipeline and their purpose.*

Name	Description
dplyr	The "grammar" of data
mirt	Psychometrics
stringr	Processing strings
RMarkdown	Reporting and generation of HTML documents
openxlsx	Reporting and generation of Excel reports

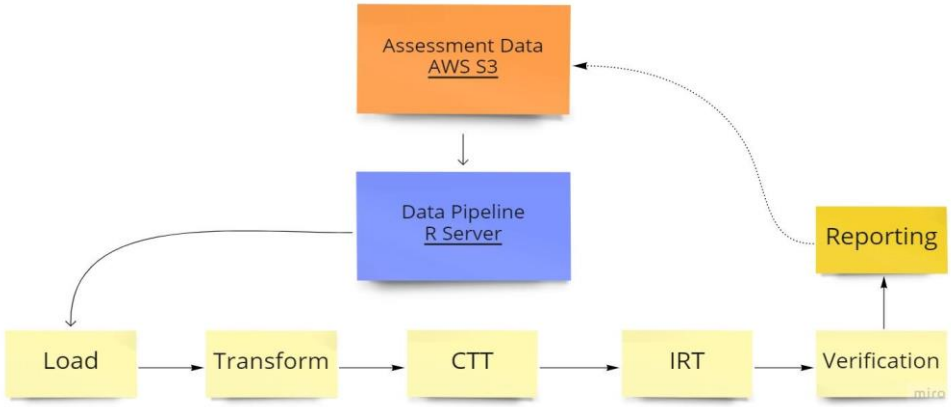
The *dplyr* package (Wickham et al., 2022) serves as the foundation for efficient data manipulation, allowing the pipeline to apply consistent and intuitive “grammar” very easily to incoming and outgoing datasets. For psychometric tasks, the project relied on the *mirt* (Chalmers, 2012) package, which offers a comprehensive suite of functions and tools

specifically tailored for psychometrics and item response theory (IRT) analysis. To handle string processing and manipulation tasks, we utilized the *stringr* package (Wickham, 2022). When it comes to generating high-quality reports, we used the powerful *RMarkdown* package (Allai*re* et al., 2022), which enables the pipeline to automatically produce dynamic HTML documents. The automated verification reports were one such document built with the package. Lastly, for generating professional-looking Excel reports, we used on the *openxlsx* package (Schauberger & Walker, 2022). This package also provided the ability to customize formatting for a specified range of cells and columns, including merging cells, bold, italics, underline, and creating borders.

2.3. Stages

The pipeline consists of several stages that must be completed successfully, from the beginning to the end, for it to run smoothly. Each stage is segmented by its purpose, with validations at each stage, so that troubleshooting is made easier. The flow of the stages in the pipeline is shown in Figure 1.

Figure 1. *A flow chart of the pipeline.*



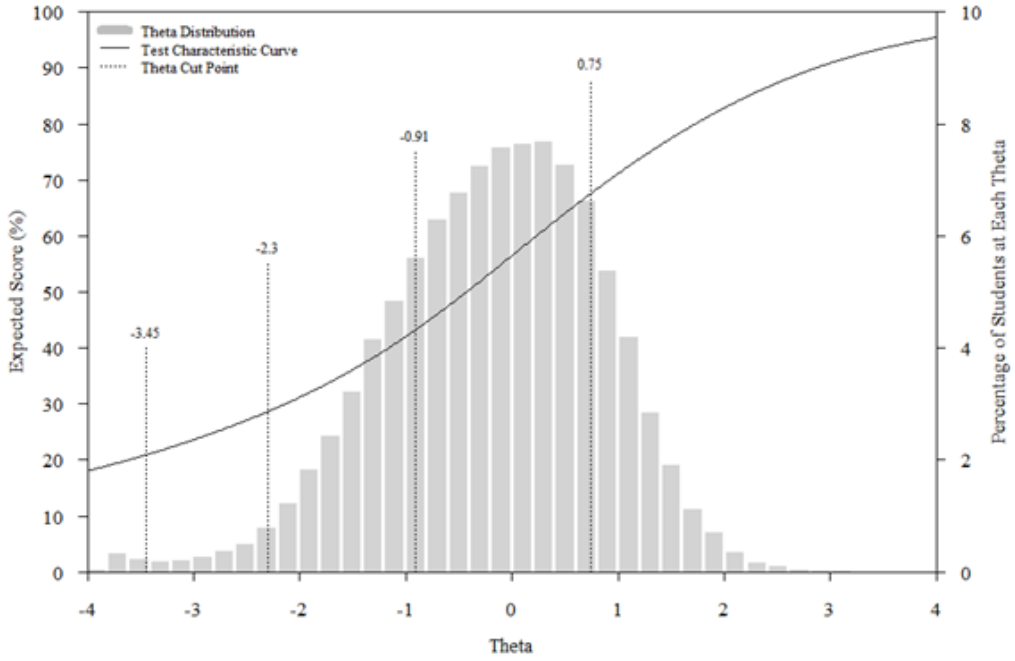
In the first stage, the standard process involves the pipeline retrieving data from Amazon S3 (Simple Storage Service), which is a cloud storage service provided by Amazon Web Services (AWS) where data is stored. If the data already exists in the local directory (the cache), the pipeline will import the data locally, saving time and processing power. We check whether the data meets the criteria for analysis and item/test specifications. At the end of this section, we create a single list of data frames that will be used in the next stages. During the transform stage, the pipeline will execute a series of processes including the application of business rules (including handling student exceptions and prorating), pivoting tables from long to wide format (for later use in IRT), and some aggregations for pre-analysis. Any kind of data cleaning is applied at this stage as well, which includes the filtering of invalid data. This stage ensures that the data frames are prepared for various types of analysis in subsequent stages.

In the CTT stage, the pipeline calculates item statistics (e.g., p, pbis, cbis) based on the CTT framework and conducts distractor analysis. These reports also include flags indicating if an item is too easy or too hard. Users have the flexibility to increase the number of flags or modify their values within the pipeline, not only in this particular section but also in other sections as per their decisions. The results of these sections help assessment teams and psychometricians to review item performance based on raw scores and frequency distributions (after the completion of the pipeline and the verification of results). For example, we can see how distractors or incorrect response options function in each item from distractor reports.

The pipeline moves to the IRT stage to conduct IRT-based analysis, including generating starting values, item calibration, equating (if necessary), and scoring (estimating thetas). In this

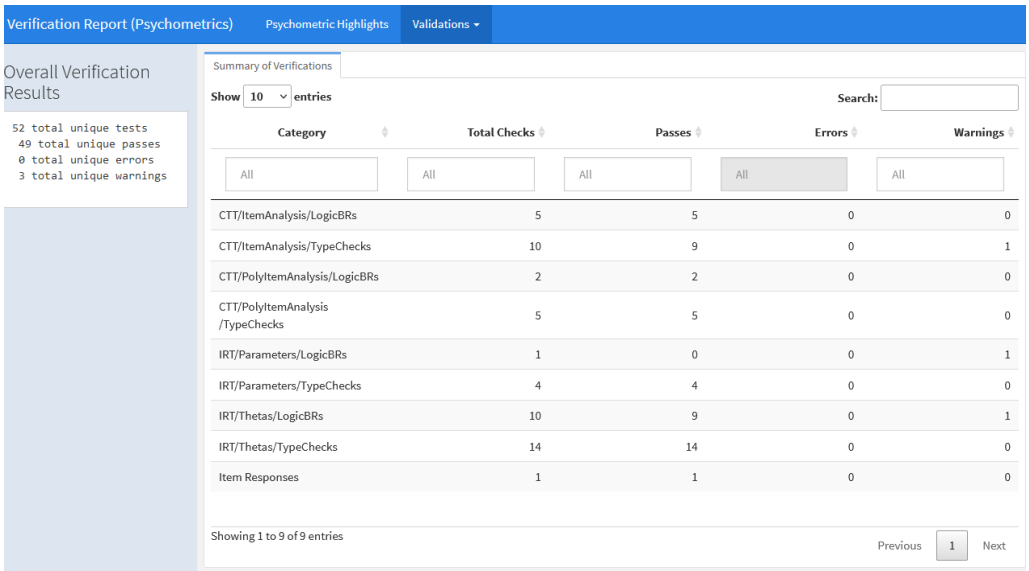
later stage, the pipeline estimates student abilities/thetas and assigns proficiency levels based on the cut scores that can be modified by the user. The pipeline produces all item- and test-level plots (see Figure 2) based on the IRT-based framework to examine individual items visually.

Figure 2. *Sample plot based on IRT framework.*



Following this stage, the pipeline generates a comprehensive report to verify the data and results prior to their reporting or publication. To ensure the accuracy and validity of the data provided to external services, the psychometrics pipeline has incorporated a robust set of data quality tests. These tests encompass the entire range of the datasets used, including both common tests applicable to all assessments and specific tests tailored to particular assessments. The tests include verifying the data format, structure of reports, constraints (nullable fields, primary and foreign keys), and business logic (consistency of the statistics reported both within a report and across several reports). The pipeline then generates an HTML report (see Figure 3) based on the outcomes of the data quality testing conducted at the conclusion of the analysis, ensuring that stakeholders and users receive accurate data.

Figure 3. *A sample page of a verification report.*



In the final stage, the pipeline generates various reports to psychometricians and content experts/assessment teams and generates database exports that are specifically designed for efficient and streamlined integration into the database, facilitating smooth and effective data transfer. Figure 4 shows examples of CTT and IRT reports.

Figure 4. *Sample pages of a CTT (below) and an IRT report (above).*

	A	B	C	D	F	G	H	I	J	K
1	label	id	ItemType	skill_category	a	b1	b2	b3	b	g
2	label.item1	item1	OR	A	1	-3.6869	-2.18543	0.35891	-1.83780667	
3	label.item2	item2	OR	A	1	-6.65895137	-2.2109488		-4.43495009	
4	label.item3	item3	MC	A	1				-1.84204	0.2
5	label.item4	item4	MC	A	1				-2.29955	0.2
6	label.item5	item5	OR	B	1	-3.19247701	-2.56037097		-2.87642399	
7	label.item6	item6	MC	B	1				-0.96892416	0.2
8	label.item7	item7	OR	C	1	-4.24080499	-2.44088468		-3.34084484	
9	label.item8	item8	MC	C	1				-0.04843	0.2
10	label.item9	item9	MC	A	1				-1.73279367	0.2
11	label.item10	item10	OR	A	1	-2.4403916	-2.69874177		-2.56956669	

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	item_label	item_id	skill	lang	type	max_	num_responded	num_nr	item_mean	Test_Score_Means	p	pbis	cpbis	tooEasy	tooDifficult
2	label.item1	item1	A	en	MC	1	35463	20	0.958	55.683	0.958	0.345	0.324	TRUE	
3	label.item2	item2	B	en	MC	1	35460	23	0.809	56.845	0.809	0.419	0.381		
4	label.item3	item3	A	en	MC	1	35457	26	0.862	56.104	0.862	0.299	0.263		
5	label.item4	item4	B	en	OR	2	35450	33	0.873	56.215	0.873	0.345	0.311		
6	label.item5	item5	B	en	OR	3	35461	22	0.725	56.324	0.725	0.235	0.186		
7	label.item6	item6	C	en	OR	3	35454	29	2.43	56.039	0.81	0.461	0.359		
8	label.item7	item7	C	en	MC	1	64264	94	0.785	58.04	0.785	0.415	0.376		
9	label.item8	item8	A	en	MC	1	10507	8	0.934	55.099	0.934	0.321	0.295	TRUE	
10	label.item9	item9	A	en	MC	1	35463	20	0.958	55.683	0.958	0.345	0.324	TRUE	
11	label.item10	item10	A	en	MC	1	35460	23	0.809	56.845	0.809	0.419	0.381		
12															

Most of the stages in the pipeline are required, but some can be skipped. For example, the CTT stage is semi-required, as most of the analysis is not required to move on to IRT. On the other hand, the verifications stage is required to be performed and confirm the data quality is clean before the pipeline can begin the reporting stage. Lastly, it is worth noting that the pipeline can be utilized either locally or online, offering flexibility in its usage.

2.4. Architecture

The RStudio Server application was installed on AWS to provide an IDE from which to log in, view the code, make changes, and run the pipeline process (end to end). We also had the capability of synchronizing accounts to provide psychometricians with the same version of the code and data. Version control (with git) was implemented using multiple branches (for production, staging, and development) in order to ensure the ability to track changes, revert changes, and to stabilize the version of the code used to produce the results.

2.5. Performance Tuning

We carried out performance tuning and identified three main areas of improvement in the pipeline. Improving the storage aspect of the pipeline was the initial focus, and it proved to be a relatively straightforward task. To enhance efficiency, a local caching system was implemented, which allows data from S3 to be stored directly in the local directory of the RStudio account. Furthermore, we underwent a thorough review of the data model, enabling us to identify and exclude unnecessary fields and tables that were not relevant to the psychometrics pipeline. As a result, these redundant components were not stored locally, optimizing storage utilization and improving overall performance.

Next, the focus shifted toward optimizing memory usage in the pipeline. This involved reducing the size of tables that contained excessive or redundant data. For instance, one particular results

dataset did not require registration data until later stages of the pipeline when specific functions were invoked. To address this, objects or datasets were loaded or called only when necessary or as closely as possible to their required usage. Additionally, we proactively employed the rm() function to remove specific objects or a list of objects from memory, and the gc() function was utilized to enforce garbage collection in R. We implemented these to remove objects as soon as they were no longer needed. These measures effectively managed memory allocation and enhanced the efficiency of the pipeline's memory utilization. It is important to acknowledge that R exhibits a tendency to consume significant memory resources and may, at times, retain memory allocation without releasing it to the operating system until explicitly required (Morandant et al., 2012).

Lastly, the most significant improvement was achieved in terms of the pipeline's execution speed. In the initial iteration, the pipeline was constructed using base R functions, resulting in a relatively slow overall performance. Using the *dplyr* package (Wickham et al., 2022) (and its *tidyr* family of packages [Wickham & Girlich, 2022]) and the data.table (Dowle & Srinivasan, 2023) resulted in speed-ups in the range of 10-100 times faster, with some calculations taking minutes instead of hours.

2.6. Integration with a Larger Administration Environment

Various options for a user interface were considered that would avoid the complexity from having to access a large codebase in R. One such consideration, the Shiny package, provides a high-level package of modules from which to build a user interface in the R language itself, rather than using Javascript, HTML, and CSS. Projects using Shiny as a frontend are generally well suited for an isolated environment where the user uploads a file and analyzes the data, with user-friendly controls. However, given its availability, the already developed and available web portal was chosen as the interface with which to interact with the psychometrics pipeline.

The final stage of pipeline development involved its full automation as well as its activation from the web portal, which was accessible by various stakeholders. This would provide the ability for the psychometricians themselves to run the entire psychometric reporting process independently, without requiring any knowledge of the codebase itself or its configuration files. Accordingly, the pipeline code was modified to be initiated with Javascript, which itself would be activated based on user input coming from the web view (including which assessment and batch of data). The web view was also modified to provide a modifiable view of the configurations used in the pipeline (cut scores, items excluded, etc.). The version of the code used, time taken, and version of the underlying data would all be automatically recorded. With the above implemented, it was possible for psychometricians to log in to the web site, choose or edit the analysis configurations, run the psychometrics for a particular assessment, and have the results delivered in a data package all from the same portal.

3. FINDINGS

The psychometrics pipeline implemented for this project took a holistic approach to data processing. It was designed to be capable of integrating with various external sources of data, including databases and data lakes. It was further able to carry the data from import to transformation, to analysis, verification, and reporting without intervention on behalf of the user. We were able to integrate this end-to-end psychometric data pipeline into a larger ecosystem that includes the registration, testing, and reporting for an assessment administration. This is crucial as organizations are looking for ways to modernize their entire administration process, and that includes the statistical analysis and reporting thereof.

The pipeline successfully conducted CTT and IRT analyses, and the results were verified by multiple independent psychometricians. A key feature of the pipeline was the generation of analysis flags and highlighting of results that required the attention of psychometricians. These

flags proved invaluable for various teams as they helped to identify and address issues promptly, enabling operational improvements. Furthermore, the pipeline facilitated psychometric work before the main administration of assessments, allowing for item piloting and review of item changes. The psychometric-related sections of the pipeline were designed to accommodate different IRT models and mixed-format test designs. While primarily utilizing the *mirt* package, the pipeline remains flexible and open to incorporating other packages, offering the capability to perform a diverse range of psychometric analyses through coding and facilitating cross-validation of results. Notably, an extension was added to the pipeline to automate test form generation based on available item banks. This extension underwent rigorous testing and successfully generated parallel test forms. Another significant extension involved running simulation studies to test various test criteria, providing valuable insights to assessment teams and test designers. The pipeline efficiently executed these simulation studies, further enhancing its capabilities and utility in the assessment process.

We implemented a strategy of self-verification within the scope of the pipeline. Requirements and constraints were understood, and a suite of data quality tests was built accordingly. These tests enabled us to perform thorough testing on the psychometric results produced by the pipeline. In this fashion, every aspect of the data could be tested, including data types, data format, data length, and business rule constraints. As the number of tests increased, the need for a visualization of the results began to be apparent. We tapped into the power of RMarkdown, which enabled us to provide fully automated reports in the style of a flexible web dashboard. This also provided the ability to more easily share and report on data quality results with stakeholders, leading to increased transparency, trust, and oversight.

Overall, we proved that R could be used to integrate with a separate system utilizing a different language and server, providing a compatible external process. Furthermore, the R packages used were overall successful in meeting our requirements. While many software packages provide a "black box" situation, we were able to dig deeper into the code used for important packages such as *mirt*, allowing us to vet the processes underneath. Such transparency and control were instrumental in ensuring the reliability and validity of our psychometric analyses. We were also able to fine-tune the performance of our R-based processes, providing a much more rapid deployment of results.

4. DISCUSSION and CONCLUSION

Modern approaches to assessment have created new requirements that are now being supported by technological innovation (Moncaleano & Russell, 2018). As the industry is modernizing, analysis is following suit. Pushed forward by provincial, national level and international testing, the industry is also beginning to adopt new approaches to handle the incoming large-scale data (Rutkowski et al., 2010). This paper presented a comprehensive solution for end-to-end data processing in large-scale assessments, addressing a significant gap in the field. Our data pipeline offers numerous benefits for practitioners, psychometricians, educators, and researchers involved in testing. It has demonstrated the ability to handle large databases, minimizing human error by automating manual processes, enabling the replication of complex workloads, ensuring high-quality outputs, and reducing overall costs associated with psychometric analysis of large-scale assessment data. By following our approach, testing organizations can enhance automation, ensure quality assurance, and achieve greater efficiency in their own large-scale assessments.

This project provided important further developments on the topic of psychometrics, data processing, administration, reporting, and the combination thereof. We also learned several lessons in developing this project. One, an understanding of the requirements and constraints of the data analysis is fundamental. We should also draw attention to the importance of having

a clear vision for the overall architecture of the pipeline. Early implementations led to costly duplication of development, human errors, and inefficiencies in running the pipeline.

We also proved that R can be a flexible and powerful tool for constructing an end-to-end data pipeline. Python is frequently used for these purposes (Weber, 2020), but we accomplished a standardized data pipeline while using the strengths of the R language. Other languages, such as Python or Julia, were not needed to fulfill the requirements for the import, transformation, analysis, and export of data. However, in the future, it would be recommended to investigate a mixed-language approach. Given Julia has an advantage in speed and memory efficiency (Dogaru & Dogaru, 2015), the language could be used for the heavy lifting by pulling and transforming data, leaving R to do the analysis and reporting.

Further embracing technological advances in recent decades would also have been beneficial in this project. Version control, at first, was quite basic, which led to the implementation of branches later in the project. As well, containerizing through Docker (Merkel, 2014) would improve the portability of the project. Docker would encapsulate the entire environment and automate all the steps it takes to build the technology architecture, installation of packages and software, and possible simulation of datasets. This would provide the ability to use the project in various mirror and user acceptance testing environments with little to no error or additional work involved in setup (Azab, 2017).

Continuous integration and continuous deployment (CI/CD) pipelines would enhance quality assurance, ensuring that updates to the code are properly tested before being deployed into production (where official results are produced). Integrated with the git repository, these CI/CD pipelines can automatically test changes made to the code before deployment, integrating unit testing and sample data into code updates. Linting packages would provide additional oversight on code syntax and style during any further development.

While there were successes throughout this project, there are some key areas that deserve further research. We noted that aspects of the integration could be improved upon. One method of enhancing the integration with the web view portal would be to transform the R code into a fully-fledged REST (Representational State Transfer) API (Application Programming Interface), using the *plumber* package (Schloerke & Allen 2023). The pipeline was integrated as a sub-process that is (except for a few configuration options) independent of the parent process that called it. An API structure would allow the pipeline to receive requests in a standardized format (using GET requests) and return data in any number of formats (csv, JSON, etc.) directly to the caller process. This would facilitate a more customized activation of the pipeline, calling certain functions and not others (running the CTT and not the IRT). This would also allow the pipeline to run asynchronously, even enabling multiple psychometrics runs at the same time.

While developing the pipeline before and during the administration windows, we found that there was a need for the large-scale generation of student data that would match the constraints of valid psychometric analysis. To this end, the simulation of student data would be an improvement to the early testing of the pipeline but also a move forward in the portability of the pipeline. As such, the pipeline could be instantiated on a fresh server, generate simulated data, and run the analysis to show that the pipeline is functioning correctly.

Lastly, the project itself was custom-built from the ground up, rather than utilizing a pre-built pipeline or finished application software like the *ShinyItemAnalysis* package (Martinková & Drabinová, 2018). Although this increased the workload, it also provided the opportunity to construct a more adaptable and customizable codebase that provides much greater functionality, tailored to the needs of each individual client. This provides a greater ability to continue to extend the project into the future, with further functionality.

Acknowledgments

This study's initial results were previously presented at the IAEA 2022 Annual Conference, held from October 2nd to 7th, 2022, in Mexico City, Mexico.

Declaration of Conflicting Interests and Ethics

The authors declare 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 authors.

H.C.B discloses that she was employed as a psychometrician at Vretta Inc. while working on and completing this paper and is currently affiliated with Northern Alberta Institute of Technology.

Authorship Contribution Statement

Ryan Schwarz: Conception, Design, Supervision, Materials, Data collection and Processing, Analysis, Literature Review, Writing. **H. Cigdem Bulut:** Conception, Design, Supervision, Materials, Data collection and Processing, Analysis, Literature Review, Writing. **Charles Anifowose:** Conception, Design, Supervision, Materials, Critical Review.

Orcid

Ryan Schwarz https://orcid.org/0009-0004-5867-3176
H. Cigdem Bulut https://orcid.org/0000-0003-2585-3686
Charles Anifowose https://orcid.org/0009-0006-2524-9613

REFERENCES

Addey, C., & Sellar, S. (2018). Why do countries participate in PISA? Understanding the role of international large-scale assessments in global education policy. In A. Verger, H.K. Altinyelken, & M. Novelli (Eds.), *Global education policy and international development: New agendas, issues and policies* (3rd ed., pp. 97–117). Bloomsbury Publishing.

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., ... & Iannone, R. (2022). *rmarkdown: Dynamic Documents for R*. R package version, 1(11).

Ansari, G.A., Parvez, M.T., & Al Khalifah, A. (2017). Cross-organizational information systems: A case for educational data mining. *International Journal of Advanced Computer Science and Applications*, *8*(11), 170-175. http://dx.doi.org/10.14569/IJACSA.2017.081122

Azab, A. (2017, April). Enabling docker containers for high-performance and many-task computing. In *2017 ieee international conference on cloud engineering (ic2e)* (pp. 279-285). IEEE.

Bezanson, J., Karpinski, S., Shah, V.B., & Edelman, A. (2012). *Julia: A fast dynamic language for technical computing.* ArXiv Preprint ArXiv:1209.5145.

Bertolini, R., Finch, S.J., & Nehm, R.H. (2021). Enhancing data pipelines for forecasting student performance: Integrating feature selection with cross-validation. *International Journal of Educational Technology in Higher Education*, *18*(1), 1-23. https://doi.org/10.1186/s41239-021-00279-6

Bertolini, R., Finch, S.J., & Nehm, R.H. (2022). Quantifying variability in predictions of student performance: Examining the impact of bootstrap resampling in data pipelines. *Computers and Education: Artificial Intelligence*, 3, 100067. https://doi.org/10.1016/j.caeai.2022.100067

Bryant, W. (2019). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research, and Evaluation*, *22*(1), 1. https://doi.org/10.7275/70yb-dj34

Camara, W.J., & Harris, D.J. (2020). Impact of technology, digital devices, and test timing on score comparability. In M.J. Margolis, R.A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 104-121). Routledge.

Chalmers. R.P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Croudace, T., Ploubidis, G., & Abbott, R. (2005). BILOG-MG, MULTILOG, PARSCALE and TESTFACT. *British Journal of Mathematical & Statistical Psychology, 58*(1), 193. https://doi.org/10.1348/000711005X37529

Desjardins, C.D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press.

Dogaru, I., & Dogaru, R. (2015, May). Using Python and Julia for efficient implementation of natural computing and complexity related algorithms. In *2015 20th International Conference on Control Systems and Computer Science* (pp. 599-604). IEEE.

Dowle, M., & Srinivasan, A. (2023). *data.table: Extension of 'data.frame'*. https://r-datatable.com,https://Rdatatable.gitlab.io/data.table.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Scientific Software International.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Erlbaum.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Kamens, D.H., & McNeely, C.L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative education review, 54*(1), 5-25. https://doi.org/10.1086/648471

Goodman, D.P., & Hambleton, R.K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145-220. https://doi.org/10.1207/s15324818ame1702_3

Liu, O.L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M.C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice, 33*(2), 19-28. https://doi.org/10.1111/emip.12028

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Addison Wesley, Reading, MA.

Lynch, S. (2022). Adapting paper-based tests for computer administration: Lessons learned from 30 years of mode effects studies in education. *Practical Assessment, Research, and Evaluation, 27*(1), 22.

IBM (2020). *IBM SPSS Statistics for Windows*, Version 27.0. IBM Corp.

Martinková, P., & Drabinová, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *R Journal, 10*(2), 503-515.

Merkel, D. (2014). Docker: Lightweight linux containers for consistent development and deployment. *Linux Journal, 2014*(239), 2.

Microsoft Corporation. (2018). *Microsoft Excel.* Retrieved from https://office.microsoft.com/excel

Moncaleano, S., & Russell, M. (2018). A historical analysis of technological advances to educational testing: A drive for efficiency and the interplay with validity. *Journal of Applied Testing Technology, 19*(1), 1–19.

Morandat, F., Hill, B., Osvald, L., & Vitek, J. (2012). Evaluating the design of the R language: Objects and functions for data analysis. In *ECOOP 2012–Object-Oriented Programming: 26th European Conference, Beijing, China, June 11-16, 2012. Proceedings 26* (pp. 104-131). Springer Berlin Heidelberg.

Muraki, E., & Bock, R.D. (2003). *PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales* [Computer software]. Scientific Software International, Inc.

Oranje, A., & Kolstad, A. (2019). Research on psychometric modeling, analysis, and reporting of the national assessment of educational progress. *Journal of Educational and Behavioral Statistics, 44*(6), 648-670. https://doi.org/10.3102/1076998619867105

R Core Team (2022). *R: Language and environment for statistical computing*. (Version 4.2.1) [Computer software]. Retrieved from https://cran.r-project.org.

Reise, S.P., Ainsworth, A.T., & Haviland, M.G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current directions in psychological science, 14*(2), 95-101.

Rupp, A.A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing, 3*(4), 365-384. https://doi.org/10.1207/S15327574IJT0304_5

Russell, M. (2016). A framework for examining the utility of technology-enhanced items. *Journal of Applied Testing Technology*, *17*(1), 20-32.

Rutkowski, L., Gonzalez, E., Joncas, M., & Von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142-151. https://doi.org/10.3102/0013189X10363170

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing" intermediate constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning and Assessment*, *4*(6).

Schauberger, P., & Walker, A (2022). *openxlsx: Read, Write and Edit xlsx Files*. https://ycphs.github.io/openxlsx/index.html, https://github.com/ycphs/openxlsx

Schleiss, J., Günther, K., & Stober, S. (2022). Protecting student data in ML Pipelines: An overview of privacy-preserving ML. In *International Conference on Artificial Intelligence in Education* (pp. 532-536). Springer, Cham.

Schloerke, B., & Allen, J. (2023). *plumber: An API Generator for R*. https://www.rplumber.io, https://github.com/rstudio/plumber

Schumacker, R. (2019). Psychometric packages in R. *Measurement: Interdisciplinary Research and Perspectives, 17*(2), 106-112. https://doi.org/10.1080/15366367.2018.1544434

Skiena, S.S. (2017). *The data science design manual*. Springer.

Sung, K.H., Noh, E.H., & Chon, K.H. (2017). Multivariate generalizability analysis of automated scoring for short answer items of social studies in large-scale assessment. *Asia Pacific Education Review*, *18*, 425-437. https://doi.org/10.1007/s12564-017-9498-1

Thissen, D., Chen, W-H, & Bock, R.D. (2003). *MULTILOG 7 for Windows: Multiple category item analysis and test scoring using item response theory* [Computer software]. Scientific Software International, Inc.

Van Rossum, G., & Drake Jr, F.L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Volante, L., & Ben Jaafar, S. (2008). Educational assessment in Canada. *Assessment in Education: Principles, Policy & Practice, 15*(2), 201-210. https://doi.org/10.1080/09695940802164226

Weber, B.G. (2020). *Data science in production: Building scalable model pipelines with Python*. CreateSpace Independent Publishing.

Wickham, H. (2022). *stringr: Simple, consistent wrappers for common string operations.* https://stringr.tidyverse.org.

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *dplyr: A grammar of data manipulation*. Retrieved from https://dplyr.tidyverse.org.

Wickham, H., & Girlich, M. (2022). *tidyr: Tidy messy data.* https://tidyr.tidyverse.org

Wise, S.L. (2018). Computer-based testing. In *the SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 341–344). SAGE Publications, Inc.

Ysseldyke, J., & Nelson, J.R. (2002). Reporting results of student performance on large-scale assessments. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*. (pp. 467-483). Routledge

Zenisky, A.L., & Sireci, S.G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337-362. https://doi.org/10.1207/S15324818AME1504_02

International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 132–148
https://doi.org/10.21449/ijate.1359348
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Automatic item generation for non-verbal reasoning items

Ayfer Sayin[1,*], **Sabiha Bozdag**[1], **Mark J. Gierl**[2]

[1]Gazi University, Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[2]University of Alberta, Faculty of Education, Department of Educational Psychology, Alberta, Canada

ARTICLE HISTORY

Received: Sep. 13, 2023
Revised: Oct. 29, 2023
Accepted: Oct. 31, 2023

Keywords:
Automatic item generation,
Non-verbal items,
Visual reasoning test,
BİLSEM.

Abstract: The purpose of this study is to generate non-verbal items for a visual reasoning test using templated-based automatic item generation (AIG). The fundamental research method involved following the three stages of template-based AIG. An item from the 2016 4th-grade entrance exam of the Science and Art Center (known as BİLSEM) was chosen as the parent item. A cognitive model and an item model were developed for non-verbal reasoning. Then, the items were generated using computer algorithms. For the first item model, 112 items were generated, and for the second item model, 1728 items were produced. The items were evaluated based on subject matter experts (SMEs). The SMEs indicated that the items met the criteria of one right answer, single content and behavior, not trivial content, and homogeneous choices. Additionally, SMEs' opinions determined that the items have varying item difficulty. The results obtained demonstrate the feasibility of AIG for creating an extensive item repository consisting of non-verbal visual reasoning items.

1. INTRODUCTION

Computer-based testing (CBT) presents several advantages, including paperless administration, flexible scheduling, and a diverse range of item types. However, CBT encounters challenges in developing continuous, content-specific items, relying on traditional item development approaches that involve experts in writing, editing, and reviewing items. To address this limitation, automatic item generation (AIG) streamlines the process through a structured workflow, ensuring a consistent supply of new, high-quality items for CBT.

The inception of AIG traces back to Bormuth's 1970 concept, which aimed to generate test items representing the intended learning outcomes (Gierl & Haladyna, 2012, p. 14). Items crafted by experts are often deemed subjective, as they reflect the experiences and personal skills of these experts. In response, Bormuth proposed automating the item writing process to eliminate subjectivity. He posited that two test developers employing the same content and item features should be capable of producing similar high-quality items (Gierl & Haladyna, 2012). AIG integrates this perspective with computer technology, marking a pioneering research field that amalgamates cognitive and psychological theories within a digital framework to generate assessment tasks (Gierl et al., 2015; Ryoo et al., 2022). The overarching goal of AIG is to

*CONTACT: Ayfer Sayın ayfersayin@gazi.edu.tr Gazi University, Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

standardize test item design significantly. By removing subjectivity from the assessment process, AIG strives to manage assessments scientifically and efficiently (Gierl et al., 2015; Leighton, 2012).

1.1. Automatic Item Generation (AIG)

AIG can be defined as a method of item generation that combines content expertise and computer technology through models, enabling the rapid creation of extensive and efficient item banks (Gierl et al., 2021). Another definition characterizes AIG as an approach to item development through augmented intelligence. Augmented intelligence is an artificial intelligence domain where computer systems model and replicate human cognitive abilities to enhance task performance (Gierl et al., 2021). The general operation of AIG necessitates the convergence of the cognitive processes shaped by subject matter experts' (SMEs) experiences and the processing power or intelligence of modern computational systems. If we conceive intelligence in the broadest sense as "problem-solving ability," AIG, with its ability to generate a vast item pool with a limited number of SMEs, demonstrates significant problem-solving capacity. AIG is based on two approaches: template-based and artificial intelligence-based (Shin, 2021). Non-verbal reasoning items were developed in the current study using template-based AIG. Template-based AIG involves a three-step standardized process. This process is explained as follows (Gierl & Lai, 2013):

- Cognitive model development: In the first step, SMEs define the content, which is referred to as the cognitive model. The cognitive model emphasizes the information, skills, and abilities required for problem solving by learners. It provides a concise depiction of subject-specific knowledge, interactions within the information, and simulates the problem-thinking/problem-solving process. It can be used not only as a template containing the relevant information, but also to provide appropriate feedback to students following exam administration.

- Item model development: In the second step, specialists decide which components of the item should be changed to establish a template for creating new items. Variables in the item model can be altered in areas such as the item's body, the question sentence, and the alternatives (right response and distractors). At this point, auxiliary elements such as photos, tables, graphs, and diagrams, as well as random variables that can be changed but are not required to answer the problem, can be introduced to the item model.

- Generating items using computer technology The content from the cognitive model is placed into the item model developed in the second phase using computer-based algorithms in the third step. In this step, computer algorithms generate objects based on the rules and limits established by SMEs. AIG has developed a variety of software, the majority of which is not open source.

Template-based AIG can be defined as generating extensive and efficient item pools by encoding content derived from the cognitive model into the item model using computer algorithms (Gierl et al., 2013). By following the three stages, AIG allows for the creation of heterogeneous item pools with similar or different item difficulties. In essence, AIG has two primary purposes: firstly, generating items with similar item difficulty with comparable psychometric properties, and secondly, constructing item pools with varying difficulty ranges (Sinharay & Johnson, 2005). This approach enables the production of items with the desired attributes and a scalable range of item difficulty.

To assess the effectiveness, performance, and suitability of AIG in response to evolving needs, it is meaningful to compare it with a conventional method, namely the traditional item writing process. From the past century to the present day, the item writing process has remained the most time-consuming and costly aspect of test development (Gierl & Haladyna, 2012). Particularly for significant tests like selection, placement, and certification, a continuous need for new items exists, leading to a demand for extensive item pools in psychometric and

educational measurements (Embretson & Yang, 2007). The traditional item writing process entails multiple steps, including item creation, item revision, and empirical testing (Embretson & Kingston, 2018). For instance, when 1000 items are required for an exam, each item must be individually authored, formatted, and developed. The elimination of items with inadequate psychometric properties at this stage further escalates costs (Arendasy & Sommer, 2012). By way of contrast, the AIG process typically commences with a well-established anchor item, which provides a robust reference point for newly generated items (Embretson & Yang, 2007). This valid anchor item contributes to the economic feasibility of AIG by satisfying a high item demand from a small number of SMEs. In short, while the traditional item writing approach ensures the creation of high-quality items, its time-consuming and cost-intensive nature renders it insufficient for meeting the increasing item demand (Choi & Zhang, 2019). Kosh et al. (2019) also highlighted the significant cost-saving potential of AIG. Moreover, items written through the traditional item writing process are limited and updating or modifying them poses challenges (Gierl et al., 2021). In our contemporary era where knowledge constantly evolves and updates, test developers require more flexible approaches. In such a context, AIG allows for the updating of items in the pool by making appropriate changes and adjustments to the previously developed cognitive model. It can be observed that the traditional item creation method is limited due to its repetitive stages, the inability to predict the psychometric properties of items without testing, the difficulty in updating generated items, and the challenge of constructing large item pools. Especially for non-verbal items, the creation of drawings and graphics is often integrated into the item writing process. This current study exemplifies the first research on the AIG process in Türkiye, which entails the generation of non-verbal items that can be used to assess students' visual reasoning skills.

1.2. Non-Verbal Reasonings

The concept of reasoning has been regarded as an ability within the domain of thinking skills (Mercan, 2021). Building upon this notion, reasoning can be defined as a cognitive process wherein an individual identifies patterns and relationships in a given problem, formulates rules, and solves the problem (Horn & Catell, 1966; Kurtz, et al., 1999). According to Mullis et al., (2019), reasoning encompasses skills such as analysis, generalization, synthesis, verification, and solving non-routine problems. Reasoning skills are considered fundamental cognitive competencies utilized in the process of accessing justified information (Kocagül & Çoban, 2022), or abstract methods and approaches used to acquire information and draw conclusions (Lawson, 2004). Reasoning skills are classified into three dimensions: mathematical/numerical, auditory/verbal, and visual-spatial/non-verbal reasoning skills (Lohman & Hagen, 2003; Mercan, 2021). The focus of the current study is on non-verbal reasoning skills, which aim to assess individuals' cognitive abilities in reasoning, independently of their verbal and language aptitudes (Balboni et al., 2010; DeThorne & Schaefer, 2004). Well-known non-verbal intelligence tests include the Universal Non-verbal Intelligence Test (UNIT), Raven Progressive Matrices (RPM), and Naglieri Non-verbal Ability Test (NNAT) (DeThorne & Schaefer, 2004). Furthermore, non-verbal reasoning items are integrated into other widely used intelligence scales in Türkiye. For instance, the Standford Binet Intelligence Test 5, the CAS Cognitive Assessment System Non-verbal Matrices subtest, and the perceptual reasoning subtest of the Weschler Intelligence Scale for Children, all employed in Guidance and Research Centers in Türkiye, incorporate items evaluating non-verbal reasoning capabilities (Gibbons & Warne, 2019; Kemer & Çakan, 2020; Naglieri et al., 2004; Weiss et al., 2016). Bildiren (2021) brought the National Non-verbal Cognitive Ability Test, a collection of non-verbal reasoning items, into the national literature. Similarly, non-verbal reasoning items were extensively used in the Visual-Perceptual Flexibility and Visual-Analogical Reasoning subtests of the Anadolu-Sak Scale, developed in Türkiye (Sak et al., 2019; Tamul et al., 2020).

Science and Art Centers (known as BİLSEM) entrance exams are conducted annually to assess candidates and identify exceptionally talented students in Türkiye. Gifted individuals are defined as children who exhibit high levels of intelligence, motivation, creativity, leadership capacity, or exceptional performance in specific academic fields compared to their peers (Bilgiç et al., 2017; MoNE, 2022a). Students nominated by their teachers for BİLSEM undergo a preliminary evaluation through a talent test determined by the BİLSEM committee for that year, administered via tablet computers (BİLSEM Online, 2023a; MoNE, 2022a). However, one of the fundamental challenges of computer-based tests is the risk of item exposure after the exam. Candidates who excel in the preliminary evaluation are subsequently subjected to individual assessment (MoNE, 2022b). Yet, especially for students nominated in the general aptitude field, the number of SMEs capable of administering intelligence tests in RAMs is limited. Moreover, many of the intelligence tests used in Türkiye lack alternative forms. Some of these tests are also outdated, which undermines the reliability of intelligence tests (Kurnaz & Ekici, 2020). Each of these factors poses a risk of item exposure in the BİLSEM entrance exam. Familiarity with the items by students who have accessed them beforehand can create a testing effect known as the practice effect, potentially affecting the results (Hausknecht et al., 2007). To mitigate this, computer-based test applications can develop personalized tests using different items for each individual or utilize adaptive applications. However, all these processes necessitate a broad repository of psychometrically sound items (Gierl & Lai, 2015). Template-based AIG can be used to generate non-verbal reasoning items quickly, economically and with high quality.

1.3. Present Study

Templated-based AIG has begun to spread across psychology, education, and computer science disciplines in recent times (Lai et al., 2016). In the literature, it has been observed that template-based AIG has been applied intensively in fields such as medicine (Falcão et al., 2022; Gierl & Lai, 2012) and dentistry (Lai et al., 2016); it has also been found to generate automated items in diverse disciplines like mathematics (Adji et al., 2018; Embretson & Kingston, 2018) and literature (Sayın & Gierl, 2023). Notably, the studies have identified verbal expressions and numerical values within mathematical items. However, the utilization of AIG in non-verbal reasoning items is limited. Gierl et al. (2015) employed template-based AIG to create 1,340 visual reasoning items involving finding the middle position and possessing heterogeneous item difficulty for undergraduate students. Ryoo et al. (2022) developed a cognitive ability test called "MOCA" that is compatible with the Cattell-Horn-Carroll (CHC) model, encompassing two of CHC's ten ability domains (Gf and Gv). MOCA, a two-form test, was designed for 6th to 9th-grade students. In contrast to both studies, the current research selected a sample group of 4th-grade elementary students and generated reasoning rotation (mental rotation) items by modifying the item format to assess their visual reasoning skills. This is because this study focuses on the Turkish sample. Visual reasoning items are used in the entrance test to BİLSEM, a school for gifted students in Türkiye, which includes visual reasoning items. The age group for the entrance exam is determined each year by the BİLSEM commission. However, considering that screening tests and diagnostic procedures have predominantly been administered to students in the 1st to 4th grades of primary school (e.g., MoNE, 2015; 2021; 2022a), 4th-grade students were prioritized when designing non-verbal reasoning items with AIG. Additionally, cognitive models were developed to create other visual reasoning items (e.g., matching, sequencing), and item generation was implemented based on these models in previous studies (Gierl et al., 2015; Ryoo et al., 2022). Unlike other studies, this research employs a rotation problem and scenario to assess visual reasoning.

AIG was achieved by utilizing a BİLSEM entrance exam item from 2016 as the primary item. In other words, the purpose of the study is to generate non-verbal reasoning items using template-based AIG. Item writing during the assessment and evaluation process is the costliest

and labour-intensive stage. Particularly in the context of visual reasoning, developing items that measure cognitive levels is a complex process requiring effort and attention. Generating items through template-based AIG will facilitate the rapid and cost-effective creation of an extensive item repository. The current study is important in terms of modeling a BİLSEM entrance exam item and serving as an example for widely used items. It also contributes to the literature and holds the distinction of creating an extensive item repository for non-verbal visual reasoning items, which is a first in Turkish literature.

2. METHOD

2.1. Research Design

This study was fundamental research, as it encompasses the automatic item generation of non-verbal visual reasoning items and their evaluation by SMEs' opinions. Fundamental research refers to investigations conducted to scrutinize, examine, reinforce, or establish a theory about a specific field (Karasar, 2022). The current study was conducted with the approval of the Gazi University Ethics Committee under the reference number E-77082166-604.01.02-686103, dated 22.06.2023.

2.2. Participants

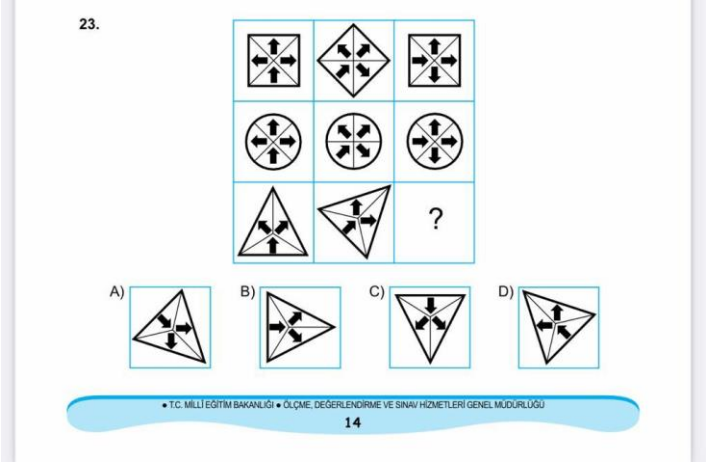
There were six participants who had previously examined BILSEM items and dealt with non-verbal items. Among the SMEs, four were female and two were male: two SMEs specializing in assessment and evaluation, one in classroom teaching, two in gifted education, and one in psychological counselling and guidance. The engagement of SMEs in the assessment and evaluation field was taken due to the test's nature and focus. In the Turkish education system, student participation in the entrance examination for a gifted education school necessitates nomination by a classroom teacher. Therefore, input from a classroom teacher was included. Given the inherent character of the test as an aptitude assessment, insights were also garnered from SMEs in the domain of gifted education. In consideration of the administration falling within the jurisdiction of psychological counsellors, the perspective of a psychological counsellor was incorporated. The SMEs, apart from classroom teachers, hold positions as university faculty members. Their professional experience varied, ranging from 5 to 17 years collectively, while their specific experience within the test development related spans 1 to 12 years.

2.3. Process

As part of the research, items were generated using AIG's three-step process. AIG generally starts with a parent item. In our study, a parent item was selected from the entrance exam for 4th-grade BİLSEM 2016 (Figure 1). BİLSEM items and data are not openly accessible. Therefore, this study concentrated on a sample exam item released by BİLSEM. While the validity evidence for the parent item could not be provided, its selection by experts in the BİLSEM commission and inclusion in the test is deemed a significant reference source.

In accordance with the parent item, the first step of AIG is the development of a cognitive model. A cognitive model represents the knowledge, skills, and abilities required to solve a specific problem within a domain. It comprehensively encompasses all the information, skills, and processes underlying test performance (Gierl & Lai, 2013). The second step of AIG focused on the development of an item model. Item models are templates that define where content needs to be placed (Gierl & Lai, 2013). The concept of an item model at AIG involves restructuring the guidelines and standards in traditional item writing using computer coding (Ryoo et al., 2022).

Figure 1. *Parent item for AIG.*



Within the current study, two item models were developed, and items were generated following these templates. The first and second steps of AIG were developed by SMEs' opinions. In the third step, computer algorithms are employed to place the content from the cognitive model into the item model, adhering to the elements and constraints defined in the cognitive model (Gierl & Lai, 2013). The prominent aspect of this process was the utilization of technology, specifically computer technology, for AIG. In our study, non-verbal visual reasoning items were generated through the utilization of the Python programming language. The codes, written in the PyCharm interface, were employed to accomplish the AIG for both developed models within the study. When the items were generated in Python, the prompt asked for the correct answers to be mixed among the options. For this reason, the correct answers were added to the bottom of each generated item and printed to an Excel file.

2.4. Data Collection Tool

The validity of the generated items was evaluated through SMEs' opinions. To facilitate this, an SME opinion form was created. A total of 20 items, 10 from each model, were presented to the SMEs for their assessment. The item-writing guidelines proposed by Haladyna et al. (2002) were utilized for the thorough examination of items by SMEs. Given the utilization of non-verbal reasoning items in our research, some criteria from the guidelines, such as 'Minimize reading, Simple vocabulary' were not used. Instead, four specific criteria were established to facilitate the comprehensive evaluation of the items: 'One right answer (Scientific Accuracy), Single content and behavior (Grade-Level Suitability Important), Not trivial content (Alignment with Purpose), homogeneous choices (Equitable challenge among distractors).' Experts assessed each item within the context of these four criteria, thus enabling the acquisition of broader and more detailed insights from the SMEs regarding the items. SMEs were requested to assess each item according to these criteria, using the following scale: 1-Accept, 2- Minor Revision, 3- Major Revision, 4- Reject. Additionally, SMEs' predictions about the difficulty of each item were obtained on a 1-5 scale, ranging from 1 as very easy to 5 as very hard.

2.5. Analysis

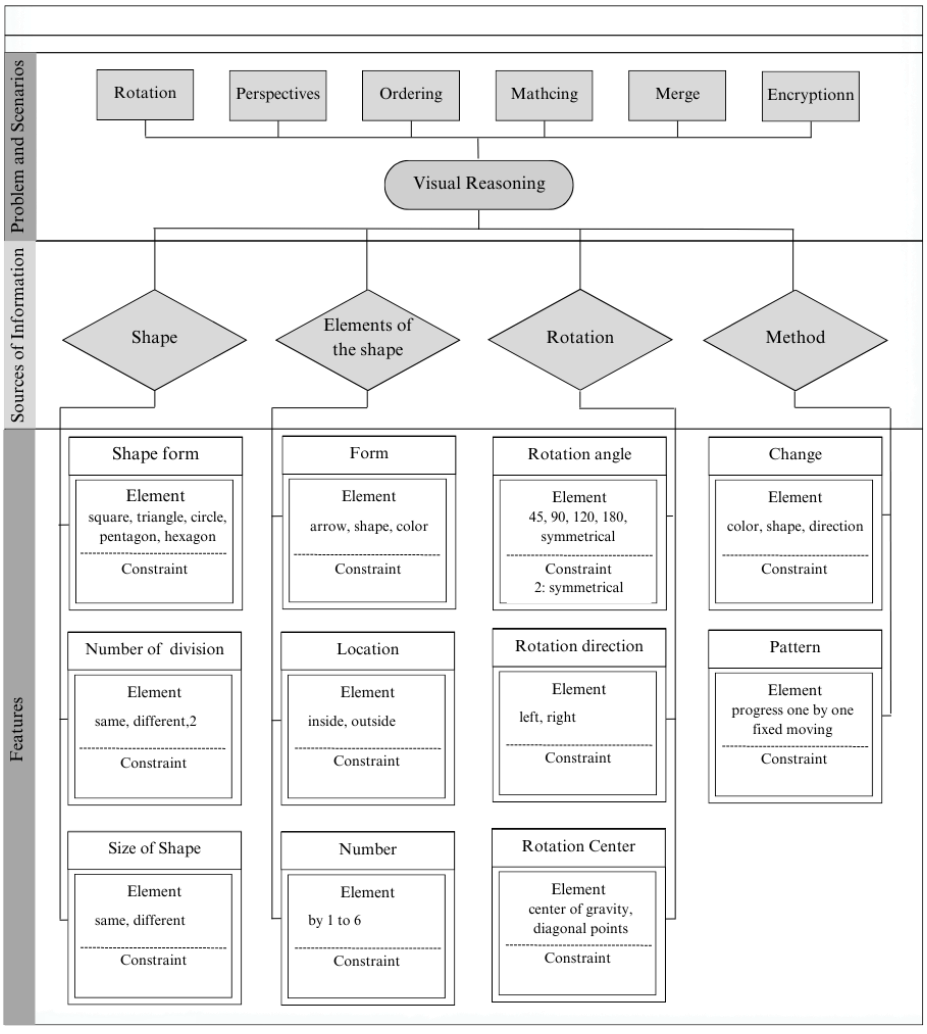
From the generated items, a random selection of 10 items was made for each model. SMEs' opinions were then collected for a total of 20 items. Frequency and percentage were calculated for the SMEs' opinions of the items.

3. RESULTS

3.1. Cognitive Model Development

The first step involved the examination of non-verbal visual reasoning items both at national and international exams, primarily focusing on BİLSEM entrance exams. The fundamental characteristics (problem and scenarios) underlying non-verbal visual reasoning items were determined as rotation, perspectives, ordering, matching, merging, and encryption. The sources of information for measuring these problems and scenarios were identified. Accordingly, the creation of distinct shapes, the incorporation of elements within or outside these shapes, and the formation of patterns through rotation and/or other methods were initially deemed essential. Once each source of information was determined, features and elements were selected. For the shape, various shapes such as square, triangle, circle, pentagon, and hexagon, among others could be chosen (as elements). These shapes could be divided into different numbers of parts, equal parts, or a fixed number like 2 to accommodate the placement of internal elements. The shapes might vary in size based on the pattern or remain consistent. Similarly, features and elements were determined for other sources of information in a manner analogous to the shape source. Afterwards, constraints were defined after the identification of elements. For instance, a triangle should be divided into 2 or 3 equal parts, while a hexagon could be divided into 6 equal parts. Nevertheless, no constraints were imposed on internal element shapes. For example, an arrow could be used in all problems and scenarios as a shape and could be incorporated within all shapes. Following these definitions, the cognitive model was developed and presented in Figure 2.

Figure 2. *A cognitive model developed for non-verbal visual reasoning items.*



In the present research, the generated items were based on the "rotation" of problems and scenarios. In this context, the developed cognitive model was structured within the framework of the Montreal Cognitive Assessment (MoCA) cognitive theory. MoCA measures visual reasoning by exploring the ability to use simulated mental images and employing the skill of rotation. In other words, it assesses students' visual reasoning skills by asking them to simulate how the movement of one shape affects another or how shapes rotate at different angles (Ryoo et al., 2022). In the current research, square, triangle, circle, and hexagon shapes were selected from the cognitive model. The square and circle were divided into four equal parts, a triangle into three equal parts, and a hexagon into six equal parts. The sizes of the shapes were constrained to the same size. Two inside elements (plus sign and square) were chosen and for these symbols, four different colors were selected: transparent, blue, green, and red. Five different angles were defined for rotation: 45, 60, 90, 120, and 180 degrees and constraints were defined for the angles according to the rhythmic logic of the shapes. For instance, the triangle shape was constrained to rotations of 60, 90, 120, and 180 degrees, while the square was constrained to rotations of 45, 90, and 180 degrees. In the first item model, rotations were carried out to the right, while in the second item model, rotations were executed to the left. All rotations were performed from the center of gravity. Elements are shown in Table 1.

3.2. Item Model Development

The parent item had a grid consisting of 3 columns * 3 rows. To showcase various item models within the study, two different item models were developed (Table 1). The question prompt was consistent for all items and was stated as "Mark the shape that should be in the place indicated by the question mark".

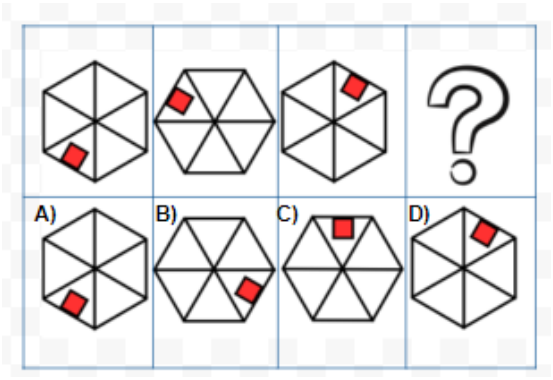
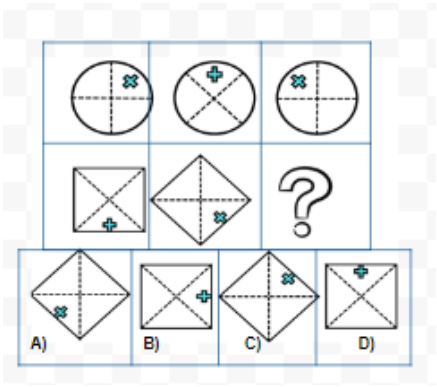
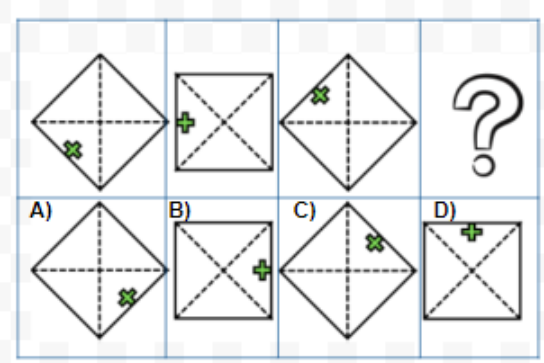
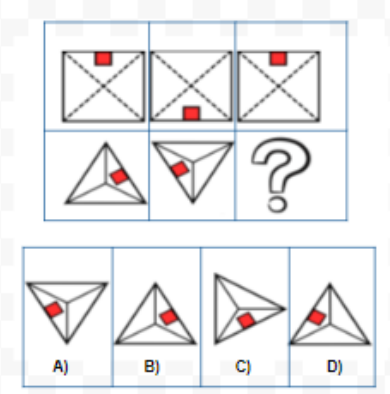
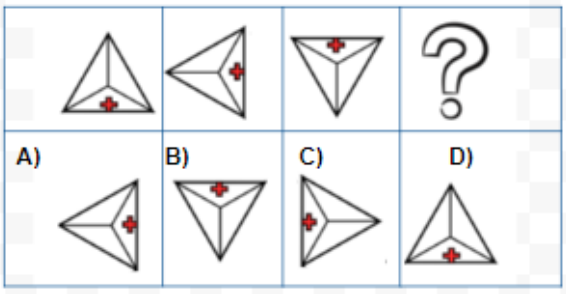
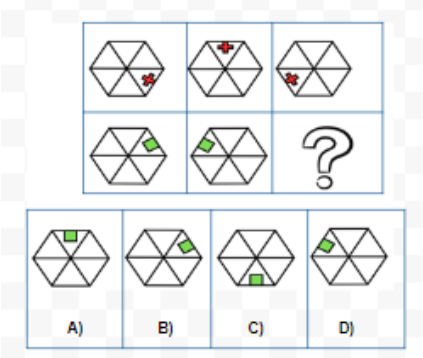
Table 1. *Item model for non-verbal reasoning items.*

Model 1	1 (column) * 4 (row) Shape x – Shape x – Shape x - ? (rotation angle and rule)
Model 2	2 (column) * 3 (row) Shape x – Shape x – Shape x (rotation angle and rule) Shape y – Shape y - ? (rotation angle and rule)
Elements	Shape_x: square, triangle, circle, hexagon Shape_y: square, triangle, circle, hexagon Rotation angle: 1. square: 45, 90, 180; 2. triangle: 60, 90, 120, 180; 3. circle: 45, 90, 180; 4. hexagon: 60, 90, 120, 180 Rotation rule: 1. right; 2. left Number of divisions: 1. square: 4, 2. triangle: 3, 3. circle: 4, 4. hexagon: 6 Internal element form: crosshair, small square Internal element color: transparent, blue, green, red
Key	Option 1, Option 2, Option 3, Option 4

3.3. Generating items using computer technology

Once the elements from the cognitive model were placed into the item model, the process of AIG for the items was initiated. At this step, Python codes were generated for each item model. 112 items from the first model and 1728 items from the second model were generated. The generated sample items are shown in Figure 3.

Figure 3. *Generated sample non-verbal reasoning items.*

Sample items from Model 1	Sample items from Model 2
1. 	1. 
Correct option: B	Correct option: B
27. 	568. 
Correct option: D	Correct option: B
51. 	1098. 
Correct option: C	Correct option: C

3.4. Review of SMEs' opinions

A random selection of 10 items was made from the generated items of each model. Opinions from 6 SMEs were gathered for the selected 20 items. The results of the SMEs' opinions were presented in Table 2 (for Model 1) and Table 3 (for Model 2). In only three items - 2, 3, and 8- minor revision suggestions had been proposed by two SMEs for Model 1. The minor revision in the 2nd item pertains to the perception that the item's difficulty was below that of the student's grade level. It had been indicated that rotating the circle 90 degrees clockwise (to the right) was considered quite manageable for 4th-grade students. The suggested minor revision for the 3rd item was about the potential challenge for students to comprehend a 60-degree rotation angle of the triangular shape. The minor revision suggested for the 8th item was oriented toward distractors. It has been recommended to insert a gap between options B and C. In the second model, for 5 items - 2, 4, 6, 7 and 8 - there exist minor revisions. For items 2 and 4, one SME has provided a visual minor revision proposal, suggesting the inclusion of gaps between distractors with rotation angles of 60 degrees each. One SME indicated the necessity for a minor revision at grade-level suitability in items 6, 7, and 8. The SME suggested that one of the distractors is relatively easy, and altering the rotational angle of this distractor had been recommended. For the items in the first model, the SMEs indicated that the difficulty ranged from very easy (1) to hard (4). Similarly, for the items in the second model, the SMEs expressed that the difficulty varied from moderately easy (2) to hard (4). In the first model, experts' opinions on the item difficulty varied between very difficult (1) and easy (5). There was no opinion suggesting that the generated items were very easy (5) in the first model. In the second model, the item difficulties were assessed by experts within the range of difficult (2) to easy (4).

Table 2. *SMEs' opinions_Model 1.*

Items	Difficulty		One right answer				Single content and behavior				Not trivial content				Choices homogeneous			
	Median	Average	Acpt.	Minor	Major	Rej	Acpt.	Minor	Major	Rej	Acpt.	Minor	Major	Rej	Acpt.	Minor	Major	Rej
I1	2	1.7	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I2	2	1.7	6	0	0	0	5	1	0	0	6	0	0	0	6	0	0	0
I3	4	3.7	6	0	0	0	4	2	0	0	6	0	0	0	6	0	0	0
I4	3	3.0	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I5	3	2.7	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I6	3	2.8	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I7	2	2.0	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I8	2	1.8	6	0	0	0	6	0	0	0	6	0	0	0	4	2	0	0
I9	1	1.7	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I10	3	2.5	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0

Table 3. *SMEs' opinions_Model 2.*

Items	Difficulty		One right answer				Single content and behavior				Not trivial content				Choices homogeneous			
	Median	Average	Acpt.	Minor	Major	Rej	Acpt.	Minor	Major	Rej	Acpt.	Minor	Major	Rej	Acpt.	Minor	Major	Rej
I1	3	3.2	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I2	3	3.3	6	0	0	0	6	0	0	0	6	0	0	0	5	1	0	0
I3	3	3.3	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I4	3	3.3	6	0	0	0	6	0	0	0	6	0	0	0	5	1	0	0
I5	4	3.5	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I6	2	2.3	6	0	0	0	5	1	0	0	6	0	0	0	5	1	0	0
I7	2	2.5	6	0	0	0	5	1	0	0	6	0	0	0	5	1	0	0
I8	3	2.8	6	0	0	0	5	1	0	0	6	0	0	0	5	1	0	0
I9	3	2.7	6	0	0	0	5	1	0	0	6	0	0	0	6	0	0	0
I10	4	3.7	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0

4. DISCUSSION and CONCLUSION

In the digital measurement and assessment era which is becoming increasingly widespread, the role and significance of visual aptitude tests are becoming even more pronounced. This is primarily due to the ease of using visual and auditory tools in digital measurement. Visual tests are used for measuring individuals' visual intelligence and problem-solving skills. They find applications in a wide range of areas, including identifying individuals with learning difficulties or gifted students, as well as in recruitment processes and career guidance (Atli, 2007; Cohen & Swerdlik, 2015). Additionally, the use of non-visual items also contributes to the validity of the test. Navigating a test in a language different from one's native tongue can pose challenges for students, particularly impacting performance. Socio-economic factors further affect students' achievement with the verbal itemsOpting for non-verbal item types to assess the special abilities of individuals from lower socio-economic backgrounds can enhance the accuracy of predictions (Bildiren et al., 2021; Lewis et al., 2007). In this case, AIG, an innovative approach to the process of creating non-verbal items, stands out. Rather than creating visuals for each item manually, utilizing computer technology can make the process more efficient and cost-effective. Therefore, AIG is used, which combines the expertise of professionals with computer technology. It has been observed in the literature that template-based AIG studies have been used in various fields such as medicine (Falcão et al., 2022; Gierl & Lai, 2012), dentistry (Lai et al., 2016), mathematics (Adji et al., 2018; Embretson & Kingston, 2018), literature (Sayin & Gierl, 2023). Also, limited studies in the existing literature, such as those by Gierl et al. (2015) and Ryoo et al. (2022), have shown that non-verbal items can be generated using AIG. Our study aimed to introduce how AIG can be used to create a comprehensive item pool focused on non-text-based items, especially for fields such as the BİLSEM entrance examination used in Türkiye (MoNE, 2022a). In this context, a cognitive model was initially developed for non-verbal visual reasoning. From the developed model, the "rotation" problem and scenario were chosen. The selected scenario was aligned with the MoCA scale, determining features and elements. Subsequently, two item models were developed. In the third step, the elements from the cognitive model were integrated into the item models using computer technology. For the first item model, 112 items were generated, while 1728 items were produced for the second item model using Python codes. This study aimed to demonstrate the applicability of non-verbal visual reasoning items with AIG. To achieve this, the range of shapes was limited by using four shapes as examples for the generated items. By increasing the number of shapes and including other elements, it is possible to create items with different similarities. Mental rotation tasks have been recognized as a measure of visuospatial ability (Cooper, 1975) and have attracted a great deal of interest in research on predicting abilities (Nolte et al., 2022). As a result, it appears as a preferred item type in BİLSEM exams. Since intelligence tests such as BNV and ASIS were integrated into the last BİLSEM entrance exams (BİLSEM Online, 2023b), the items were not opened. However, six mental rotation items were identified in the 50-item test (2016 test), which included only parent items, indicating that these items were used by changing the shape-rotation angle. This suggests that the items created in this study may find application in tests from different years.

In our study, the generated items were evaluated based on the SMEs` opinions. SMEs examined randomly selected 10 items from each model based on four different criteria and predicted the item difficulty. As a result of SMEs' opinions, it was determined that the items have varying item difficulty. This outcome was anticipated since the positions of different shapes at the same rotation angle exhibit variation. For instance, a square manifests a more pronounced 45-degree rotation angle than a circle due to its four sides. This discrepancy in rotation angle/speed is attributed to the size of the central mass (shape) and the congruence of sides and angles. Given that the main body size of a quadrilateral surpasses that of a triangle, the perceived rotation

speed is heightened (Pylyshyn, 1979). Consequently, experts rated triangle-related items as more challenging than their quadrilateral counterparts. Additionally, the obtained results suggest the feasibility of generating items by constructing item pools with varying difficulty ranges by AIG (Sinharay & Johnson, 2005). The findings indicated that item difficulty, as perceived by experts, also varied based on the models. In the parent item, showcasing the 3*3 item model, instances of the desired pattern were presented in the final line in two distinct forms. These instances eased problem-solving by providing additional information. Similarly, items generated with the 2*3 item model in this study offer more information about problem resolution compared to items created with the 1*4 item model. Because it includes two lines for the solution. This clarifies why experts considered items from the 2*3 item model easier than those from the 1*4 item model. Furthermore, the uniform used of a single rotation rule and one internal element in both item models contributed to a general evaluation of items easily. Ultimately, item difficulties varied based on the item model and the elements. It shows that introducing new cognitive features to the item model has the potential to yield more intricate items.

Generated items were appropriate for one right answer, single content and behavior, not trivial content, and choices homogeneous by SMEs. The obtained results demonstrated the applicability of AIG for a comprehensive pool of items consisting of non-verbal visual reasoning items. Throughout the process of generating non-verbal visual items using AIG, it was noticed that the role of SMEs in shaping the scientific and item model is critically crucial. The contributions of SMEs had aided in ensuring the accuracy of item content. And it showed that the innovations brought about by utilizing computer technology had shown that it could efficiently and cost-effectively create a large item pool. This technological advancement has the potential to make it more efficient and accessible. Based on the findings of the current research, we recommend the creation of a comprehensive item pool using the results obtained. This item pool can be effectively utilized in computer-based tests, offering the advantages of personalized testing, and adaptive testing, and allowing multiple test administrations within a year. These recommendations are crucial for enhancing the evaluation of student performance and supporting more effective learning processes. Furthermore, we suggest future research initiatives, such as conducting field research and exploring equivalence for test equating. These advanced studies can further optimize the process of AIG and enhance the existing knowledge base in this field. In conclusion, the current study emphasizes the significance of visual aptitude tests in meeting the demands of contemporary digital assessments and highlights the feasibility of generating such tests using AIG. By demonstrating how AIG can facilitate the creation of a comprehensive item pool, especially for assessments used in in Türkiye, the current research aims to lay the groundwork for future research and applications in the realms of education and assessment.

Limitations

Acknowledgments of people, grants, and funds should be placed in a separate section before the References. If the study has been previously presented at a conference or a scholarly meeting, it should be mentioned here. The present study focused on exploring the viability of generating non-verbal reasoning items through AIG, with item evaluation conducted based on expert opinions. For future investigations, it would be beneficial to conduct field tests on the AIG-generated test items and estimate validity evidence by analyzing the data coming from field tests. The potential for a testing effect arises when items from the same pool are employed at different times, particularly within short-term intervals. To mitigate this, diversifying the item pool by varying elements and item models could be considered. Additionally, since this study exclusively utilized rotation, it is advisable to incorporate item samples that assess other problem situations in future research endeavours.

Acknowledgments

The authors would like to thank the blind reviewers and SMEs for their useful comments and insightful suggestions.

Declaration of Conflicting Interests and Ethics

The authors declare 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 authors. **Ethics Committee Number**: Gazi University Ethics Committee, E-77082166-604.01.02-686103. 22.06.2023.

Authorship Contribution Statement

Ayfer Sayin: Design, Data Collection and/or Processing, Analysis and Interpretation, Literature Review, Writing. **Sabiha Bozdag:** Materials, Data Collection and/or Processing, Literature Review, Writing. **Mark J. Gierl:** Conception, Design, Supervision, Writing, Critical Review

Orcid

Ayfer Sayin https://orcid.org/0000-0003-1357-5674
Sabiha Bozdag https://orcid.org/0000-0002-2039-8066
Mark J. Gierl https://orcid.org/0000-0002-2653-1761

REFERENCES

Adji, T.B., Pribadi, F.S., Prabowo, H.E., Rosnawati, R., & Wijaya, A. (2018). Generating parallel mathematic items using automatic item generation. *ICEAP 2019, 1*(1), 89-93. https://doi.org/10.26499/iceap.v1i1.78

Arendasy, M.E., & Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and individual differences, 22*(1), 112-117. https://doi.org/10.1016/j.lindif.2011.11.005

Atli, S. (2007). *Matematiksel mantıksal yetenek ile ritimsel yetenek arasındaki ilişkiler* [*Relations between matematical-logical talent and rhythmic intelligence*] [Unpublished master's thesis]. Gazi University.

Balboni, G., Naglieri, J.A., & Cubelli, R. (2010). Concurrent and predictive validity of the raven progressive matrices and the Naglieri Nonverbal Ability Test. *Journal of Psychoeducational Assessment, 28*(3), 222-235. https://doi.org/10.1177/073428290934376

Bildiren, A., Bıkmaz Bilgen, Ö., & Korkmaz, M. (2021). National non-verbal cognitive ability test (BNV) development study. *SAGE Open, 11*(3). https://doi.org/10.1177/2158244021104694

Bilgiç, N., Taştan, A., Kurukaya, G., Kaya, K., Avanoğlu, O., ve Topal, T. (2017). *Özel yetenekli bireylerin eğitimi strateji ve uygulama kılavuzu [Education of specially gifted individuals' strategy and implementation guide].* MEB Özel Eğitim ve Rehberlik Hizmetleri Genel Müdürlüğü. https://orgm.meb.gov.tr/meb_iys_dosyalar/2013_11/25034903_zelyeteneklibireylerineitimistratejiveuygulamaklavuzu.pdf

BİLSEM Online (2023a). *Sıkça sorulan sorular: BİSEM sınav soruları yeteneğe göre değişir mi? [Frequently asked questions: Do BİLSEM exam questions vary depending on ability?].* https://www.bilsemonline.com/sss

BİLSEM Online (2023b). BNV Zeka Testi Nedir? *[What is BNV Intelligence Test?].* https://bilsemonline.com/blog/bnv-zeka-testi-nedir

Choi, J., & Zhang, X. (2019). Computerized item modeling practices using computer adaptive formative assessment automatic item generation system: A tutorial. *The Quantitative Methods for Psychology, 15*(3), 214-225. https://doi.org/10.20982/tqmp.15.3.p214

Cohen, R.J., & Swerdlik, M.E. (2015). *Psychological testing and assessment*. McGraw-Hill Education.

Cooper, L.A. (1975). Mental rotation of random two-dimensional shapes. *Cognitive Psychology, 7*(1), 20-43. https://doi.org/10.1016/0010-0285(75)90003-1

DeThorne, L.S. & Schaefer, B.A. (2004). A guide to child nonverbal IQ measures. *American Journal of Speech-Language Psychology, 13*(4), 275-290. https://doi.org/10.1044/1058-0360(2004/029)

Embretson, S.E., & Kingston, N.M. (2018). Automatic item generation: A more efficient process for developing mathematics achievement items? *Journal of Educational Measurement, 55*(1), 112-131. https://doi.org/10.1111/jedm.12166

Embretson, S., & Yang, X. (2007). 23 Automatic item generation and cognitive psychology. *Handbook of statistics, 26*, 747-768. https://doi.org/10.1016/S0169-7161(06)26023-1

Falcão, F., Costa, P., & Pêgo, J.M. (2022). Feasibility assurance: a review of automatic item generation in medical assessment. *Advances in Health Sciences Education, 27*(2), 405-425. https://doi.org/10.1007/s10459-022-10092-z

Gibbons, A., & Warne, R.T. (2019). First publication of subtests in the Stanford-Binet 5, WAIS-IV, WISC-V, and WPPSI-IV. *Intelligence, 75*, 9-18. https://doi.org/10.1016/j.intell.2019.02.005

Gierl, M.J., & Haladyna, T. (2012). *Automatic item generation: Theory and practice*. Routledge.

Gierl, M.J., & Lai, H. (2012). The role of item models in automatic item generation. *International Journal of Testing, 12*(3), 273-298. https://doi.org/10.1080/15305058.2011.635830

Gierl, M.J., & Lai, H. (2013). Instructional topics in educational measurement (ITEMS) module: Using automated processes to generate test items. *Educational Measurement: Issues and Practice, 32*(3), 36-50. https://doi.org/10.1111/emip.12018

Gierl, M.J. & Lai, H. (2016). Automatic item generation. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd edition, pp. 410-429). Routledge.

Gierl, M.J., Ball, M.M., Vele, V., & Lai, H. (2015). A method for generating nonverbal reasoning items using n-layer modeling. *In Computer Assisted Assessment. Research into E-Assessment: 18th International Conference, CAA 2015*, Zeist, The Netherlands, June 22–23, 2015. https://doi.org/10.1007/978-3-319-27704-2_2

Gierl, M., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation*. Routledge.

Haladyna, T. M., Downing, S. M. & Rodriguez, M. C. (2002) A review of multiple-choice item-writing guidelines for classroom assessment, *Applied Measurement in Education, 15*(3), 309-333, https://doi.org/10.1207/S15324818AME1503_5

Hausknecht, J.P., Halpert, J.A., Di Paolo, N.T., & Moriarty Gerrard, M.O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*(2), 373–385. https://doi.org/10.1037/0021-9010.92.2.373

Horn, J.L., & Cattell, R.B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*(5), 253-270. https://doi.org/10.1037/h0023816

Karasar, N. (2022). *Bilimsel araştırma yöntemleri (37. Basım)*. Nobel Yayıncılık.

Kemer, B., & Çakan, M. (2020). Examining the validity of the psychological scales frequently used in guidance and research centers with respect to measurement standards of validity. *Journal of Research in Education and Society, 7*(1), 323-348. https://dergipark.org.tr/en/pub/etad/issue/55359/731549

Kocagül, M., & Çoban, G.Ü. (2022). An evaluation on science teachers' scientific reasoning skills. *Cumhuriyet International Journal of Education, 11*(2), 361-373. https://doi.org/10.30703/cije.1017938

Kosh, A.E., Simpson, M.A., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2019). A cost–benefit analysis of automatic item generation. *Educational Measurement: Issues and Practice, 38*(1), 48-53. https://doi.org/10.1111/emip.12237

Kurnaz, A., & Ekici, S.G. (2020). BİLSEM tanılama sürecinde kullanılan zeka testlerinin psikolojik danışmanların ve BİLSEM öğretmenlerinin görüşlerine göre değerlendirilmesi [Evaluation of intelligence tests used in BİLSEM diagnostic process according to the opinions of psychological counselors and BİLSEM Teachers]. *Çocuk ve Medeniyet, 5*(10), 365-399. https://dergipark.org.tr/en/pub/cm/issue/59377/850922

Kurtz, K., Gentner, D., & Gunn, V. (1999). Reasoning. In B. M. Bly & D. E. Rumelhart (Eds), *Cognitive science*, pp. 145-200. California: Academic Press.

Lai, H., Gierl, M.J., Byrne, B.E., Spielman, A.I., & Waldschmidt, D.M. (2016). Three modeling applications to promote automatic item generation for examinations in dentistry. *Journal of Dental Education, 80*(3), 339-347. PMID: 26933110.

Lawson, A.E. (2004). The nature and development of scientific reasoning: A synthetic view. *International Journal of Science and Mathematics Education, 2*(3), 307-338. https://doi.org/10.1007/s10763-004-3224-2

Leighton, J.P. (2012). Learning sciences, cognitive models, and automatic item generation. In M.J. Gierl, & T.M. Haladyna (Eds), *Automatic item generation: Theory and practice*, pp. 121-135. Routledge.

Lewis, J.D., DeCamp-Fritson, S.S., Ramage, J.C., McFarland, M.A., & Archwamety, T. (2007). Selecting for ethnically diverse children who may be gifted using Raven's Standard Progressive Matrices and Naglieri Nonverbal Abilities Test. *Multicultural Education, 15*(1), 38-42.

Lohman, D.F., & Hagen, E. (2003). *Interpretive guide for teachers and counselors: cognitive abilities test Form 6-all levels*. ITASCA, Illinois: Riverside Publishing.

Mercan, Z. (2021). Studies on early childhood reasoning skills in Turkey. *Journal of Muallim Rıfat Faculty of Education, 3*(2), 104-120.

Ministry of National Education (MoNE) (2015). Bilim ve Sanat Merkezleri Yönergesi [Science and Art Centers Directive]. MoNE General Directorate of Special Education and Guidance Services. https://orgm.meb.gov.tr/meb_iys_dosyalar/2015_10/26091626_blsemkilavuz26.10.2015.pdf

Ministry of National Education (MoNE) (2021). Bilim ve Sanat Merkezleri Yönergesi [Science and Art Centers Directive]. MoNE General Directorate of Special Education and Guidance Services]. https://orgm.meb.gov.tr/meb_iys_dosyalar/2021_12/30144032_2021-2022_YILI_BILIM_VE_SANAT_MERKEZLERI_OGRENCI_TANILAMA_VE_YERLESTIRME_KILAVUZU.pdf

Ministry of National Education (MoNE) (2022a). *Bilim ve Sanat Merkezleri Yönergesi [Science and Art Centers Directive]*. MoNE General Directorate of Special Education and Guidance Services. https://orgm.meb.gov.tr/meb_iys_dosyalar/2016_10/07031350_bilsem_yonergesi.pdf

Ministry of National Education (MoNE) (2022b). *Bilim ve Sanat Merkezleri öğrenci tanılama ve yerleştirme kılavuzu [Science and Art Centers student identification and placement guide]*. MoNE General Directorate of Special Education and Guidance Services. https://orgm.meb.gov.tr/www/bilsem-ogrenci-tanilama-ve-yerlestirme-kilavuzu-yayimlandi/icerik/2154

Mullin, I.V.S., Martin, M.O., & Foy, P. (2005). *IEA's TIMSS 2003 international report on achievement in the mathematics cognitive domains*. TIMSS & PIRLS Internatonal study

Center. Lynch School of Education, Boston College. https://files.eric.ed.gov/fulltext/ED494652.pdf

Naglieri, J.A., & Ford, D.Y. (2005). Increasing minority children's representation in gifted education: A response to Lohman. *Gifted Child Quarterly*, *49*(1), 29-36. https://doi.org/10.1177/001698620504900104

Nolte, N., Schmitz, F., Fleischer, J., Bungart, M., & Leutner, D. (2022). Rotational complexity in mental rotation around cardinal and skewed rotation axes. *Intelligence*, *91*, 101626. https://doi.org/10.1016/j.intell.2022.101626

Pylyshyn, Z. W. (1979). The rate of "mental rotation" of images: A test of a holistic analogue hypothesis. *Memory & Cognition, 7*(1), 19-28. https://doi.org/10.3758/BF03196930

Ryoo, J.H., Park, S., Suh, H., Choi, J., & Kwon, J. (2022). Development of a new measure of cognitive ability using automatic item generation and its psychometric properties. *SAGE Open*, 1-13. https://doi.org/10.1177/21582440221095016

Sak, U., Sezerel, B.B., Dulger, E., Sozel, K., & Ayas, M.B. (2019). Validity of the Anadolu-Sak Intelligence Scale in the identification of gifted students. *Psychological Test and Assessment Modeling, 61*(3), 263-283. https://psycnet.apa.org/record/2020-53108-001

Sayın, A., & Gierl, M.J. (2023). Automatic item generation for online measurement and evaluation: Turkish literature items. *International Journal of Assessment Tools in Education, 10*(2), 218-231. https://doi.org/10.21449/ijate.1249297

Shin, E. (2021). *Automated item generation by combining the non-template and template-based approaches to generate reading inference test items.* [Doctoral dissertation, University of Alberta]. Education and Research Archive. https://doi.org/10.7939/r3-75wr-hc80

Sinharay, S., & Johnson, M. (2005). Analysis of data from an admissions test with item models. *ETS Research Report Series, 2005*(1), 1-32. https://files.eric.ed.gov/fulltext/EJ1111287.pdf

Tamul, Ö.F., Sezerel, B.B., Sak, U., & Karabacak, F. (2020). Social validity study of the Anadolu-SAK intelligence scale (ASIS). *PAU Journal of Education, 49*, 393-412. https://doi.org/10.9779/pauefd.575479

Weiss, L.C., Saklofske, D.H., Holdnack, J.A., & Prifitera, A. (2016). WISC-V: Advances in the assessment of intelligence. In L.G. Weiss, D.H. Saklofske, J.A. Holdnack, & A. Prifitera (Eds.), *WISC-V assessment and interpretation: Scientist-practitioner perspectives* (pp. 3–23). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-404697-9.00001-7

International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 149–163
https://doi.org/10.21449/ijate.1394194
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Language models in automated essay scoring: Insights for the Turkish language

Tahereh Firoozi[1*], **Okan Bulut**[1], **Mark J. Gierl**[1]

[1]University of Alberta, Edmonton, Alberta, Canada

ARTICLE HISTORY

Received: Nov. 22, 2023
Accepted: Dec. 17, 2023

Keywords:
Automated essay scoring,
Word embedding,
Transformers,
BERT,
Turkish AES.

Abstract: The proliferation of large language models represents a paradigm shift in the landscape of automated essay scoring (AES) systems, fundamentally elevating their accuracy and efficacy. This study presents an extensive examination of large language models, with a particular emphasis on the transformative influence of transformer-based models, such as BERT, mBERT, LaBSE, and GPT, in augmenting the accuracy of multilingual AES systems. The exploration of these advancements within the context of the Turkish language serves as a compelling illustration of the potential for harnessing large language models to elevate AES performance in in low-resource linguistic environments. Our study provides valuable insights for the ongoing discourse on the intersection of artificial intelligence and educational assessment.

1. LANGUAGE MODELS IN AUTOMATED ESSAY SCORING

Automated essay scoring (AES) is a sub-task of text classification that uses computer algorithms to score essays written by humans automatically. Machine and deep learning algorithms are often utilized to build a scoring engine that can model the scoring performance of human raters. The model is then employed to classify essays into different score classes (i.e., score categories). AES systems typically work by analyzing the text of an essay and applying a set of linguistic features to assess its quality. These features may include grammar, vocabulary, sentence structure, coherence, and the presence of relevant arguments or evidence. The AES system builds the scoring model using techniques and procedures from the fields of natural language processing (NLP) and computational linguistics where linguistic features are extracted from the instances of human-scored essays (i.e., labeled data) and turned into numerical representations that a machine or deep learning model can process. The most common NLP techniques for feature extraction include text length features, bag of words, and pre-trained large language models such as Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2018).

Text length features are simple and effective in general text analysis and AES (Fleckenstein et al., 2020; Hussein et al., 2019), as they have been widely used to evaluate essays based on their

*CONTACT: Tahereh FIROOZI ✉ tahereh.firoozi@ualberta.ca University of Alberta, Edmonton, Alberta, Canada

e-ISSN: 2148-7456 /© **IJATE** 2023

length and structure. These features include, for example, the average number of words per paragraph, or the average number of characters per word. Using text length features, AES systems compare the length of each essay to the length of the essay prompt or an ideal essay length with a high score in the training corpus. Text length features are typically used in conjunction with other syntactic properties, such as part of speech (POS) and discourse characteristics of a text, including cohesion and coherence. Statistical language models are used to analyze the syntactic properties in a text (e.g., Rodriguez et al., 2019). N-gram is an example of the statistical model that captures the likelihood of a sequence of n words occurring in a given text based on the frequency of those word sequences in a training corpus. Practically, the syntactical property of texts is assessed using the existing natural language toolkit libraries in programming languages, such as the NLTK library in Python (Bird, 2006). In addition, the linguistic and discourse characteristics of written texts in English can be assessed using text analysis tools, such as Coh-Metrix (Graesser et al., 2004), which were developed and used for the English language.

Word embeddings have become fundamental in various NLP applications, including AES. Word embedding models, such as Word2vec and GloVe embeddings, are a class of NLP techniques to represent words as dense vectors in a continuous vector space (Firoozi et al., 2022). These vectors capture hidden information about a language, like word analogies or semantics. The information can be used to examine the proximity of the semantic relationship between the word and the context. For example, in a well-trained word embedding model, the vectors for "king" and "queen" would be closer together than the vectors for "king" and "car." Calculating the proximity in the vector space allows the model to capture semantic relationships, such as analogies ("king" is to "queen" as "man" is to "woman"). This knowledge is learned in pre-trained word embedding models through unsupervised learning using large amounts of text corpus. Depending on the corpus and the learning techniques, the word embedding models capture different information in the vectors. The most popular word embedding language models are Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017) developed by Google Inc., Stanford University, and Facebook AI Research, respectively. The models were pre-trained using different corpora and learning techniques.

Word2Vec pre-trained vectors were trained on a part of the Google News dataset with about 100 billion words. Mikolov et al. (2013) proposed two model architectures, including a continuous bag of words (CBOW) and Skip-gram, for learning distributed representations of words. CBOW predicts a word in a sequence of words given the average distributed representation of all the surrounding words in the sequence. A word is predicted based on the largest semantic similarity between the word vector and the average distributed representation vector of the surrounding words in context. The Skip-gram model is similar to CBOW, but instead of predicting a word based on the surrounding sequence of words, it tries to predict the surrounding words of a given word in a sentence.

FastText (Bojanowski et al., 2017) has a similar training process (i.e., CBOW and Skip-gram) as Word2Vec. FastText differs from Word2Vec regarding the corpora used for training and word representation technique. FastText is trained on Wikipedia data in nine different languages: Arabic, Czech, German, English, Spanish, French, Italian, Romanian, Russian, and Turkish (Kuyumcu et al., 2019). FastText represents words as bags of character n-grams (subword units). In this technique, words are decomposed into character-level n-grams (e.g., "apple" -> {ap, ppl, ple}), including both prefixes and suffixes. Word representations are then generated by summing or averaging the vectors of these n-grams. The character level representation in FastText enables the model to capture morphological and semantic information even for out-of-vocabulary words.

The training process of the Glove model (Pennington et al., 2014) is different from Word2Vec. The GloVe model combines the matrix factorization methods (Cai et al., 2009) and the window-based methods to consider both the statistical and contextual information of words in calculating word vectors. Hence, GloVe learns the embeddings based on a co-occurrence matrix showing the count of the overall statistics of how often words appear together in a text based on their semantic similarity. The vector spaces of the word embedding techniques can be trained on AES datasets with different sizes to fine-tune the pre-trained parameters.

The word embedding models use Recurrent Neural Network (RNN) models for training (Liu et al., 2015). RNN models are neural network models containing a hidden layer that auto-regressively updates the conditional probability of the output vector (e.g., a word or the context of a word) given the hidden state in the next step. The RNN model updates the prediction weights based on the errors it receives in the following steps. While RNN-based models revolutionized the Google translation engines in 2016, they have two main problems. First, these models suffer from the vanishing gradient problem, making it very difficult to capture long-range dependencies within the text (Hochreiter et al., 2001). For instance, if a system is developed to predict the next word in a sentence, the network must have a better knowledge of the preceding words in the text for more accurate predictions. In RNN, the hidden weights are updated recurrently to decrease the error function. In long texts with more hidden weights at different time steps, the initial weights are multiplied by the updated weight. However, because the initial weights are small, this multiplication quickly decreases the gradient value, leading to the early termination of model training before the model can learn the whole text. This problem with sequential training was solved using a parallel structure in encoding the input sequence of different lengths (Vaswani et al., 2017).

2. TRANSFORMER_BASED LANGAUGE MODELS

Research studies on AES skyrocketed in 2018 when the transformer models (Vaswani et al., 2017) were introduced (Ramesh & Sanampudi, 2022). Transformer-based models have revolutionized the field of NLP by offering powerful tools for training language models that significantly increase the accuracy of the pre-trained models in text classification tasks, including AES (Devlin et al., 2018). Transformers are encoder-decoder-based neural networks that solve sequence processing problems by finding a mapping function from an input sequence of vectors (e.g., word or sentence) to the output sequence of vectors (e.g., essay labels). The architecture consists of an encoder and a decoder comprising multiple layers of multi-head attention-based blocks. The encoder takes the input sequence and processes it by repeatedly applying the multi-head attention block to the input sequence of tokens. The attention mechanism in transformers can capture all of the contextual information within a text to calculate the weighted sum of values for each token (e.g., words) in a sequence of input (e.g., sentence). For example, in the sequence of input = “I want to buy a car,” the representation of the fourth word “buy” depends not only on the adjacent words, including “I,” “want,” “to,” “a,” and “car” in the sequence, but also on all other words in the text. This feature allows for the modeling of global dependencies in all sequential inputs without regard to their distance in the input or output sequences. Hence, in encoding or decoding the representation of an input sequence, the attention mechanism allows transformers to learn the context of the input by parallelizing all the surrounding inputs within training examples (Wolf et al., 2020). The decoder takes the output sequence generated by the encoder and processes it by attending to the encoder's output and the previous tokens in the decoder's input sequence. This allows the decoder to generate each output token by selectively attending to different parts of the input sequence. During training, the model learns to assign appropriate weights to the tokens in each layer of the encoder and decoder in order to minimize a given loss function (e.g., cross-entropy

loss) between the predicted output sequence and the ground truth output sequence (Han et al., 2022).

Using this training approach, transformer-based language models can accurately capture the linguistic patterns in a language by learning the long-range dependencies and semantic relationships between tokens (e.g., words/subwords or sentences) in texts (Bouschery et al., 2023). Hence, by transferring the knowledge in these pre-trained language models to the targeted AES task, the accuracy of the AES systems improves significantly without using a large number of labeled essays for training. Unlike the text length and word embedding models that were language-specific and mainly developed and used for the English language, transformer-based models were also trained on languages other than English. Hence, the low-resource languages can benefit from transformer-based language models, including BERT, LaBSE, and GPT, for transfer learning in AES (Firoozi et al., 2023). Low-resource languages are less studied, less digitized, less privileged, less commonly taught, and less accessible compared to English (Cieri et al., 2016; Magueresse et al., 2019).

BERT (Devlin et al., 2018) is a transformer-based encoder model for language representation that uses a multi-head attention mechanism and a bidirectional approach to learn the contextual relations between words and sentences in a text for an accurate representation of the entire text. Multi-head attention is a mechanism for training transformers that compares each input vector with all other text vectors to consider the context in word representations. The bidirectional in BERT refers to the training process where the transformers can generate contextual embeddings based on previous and next tokens of the text (Kenton & Toutanova, 2019). BERT can be trained in different languages. BERT is trained using two approaches. The first approach is masked language modeling (MLM). MLM predicts a missing word in a sentence by randomly masking 15% of the words and running all the masked sentences through the model to predict the masked words. The second approach is next sentence prediction (NSP). NSP is predicting if one sentence naturally follows another. In NSP, the model learns to understand longer-term dependencies across sentences. Using the NSP technique allows the model to predict each two-sentence sequence that follows one another in a text. BERT learns this knowledge by receiving masked sentence embeddings concatenated in pairs as inputs during pre-training. Half of the embeddings are random, and the other half are actual sentence pairs from the pool of training data. For example, the model receives sentence A and sentence B to predict whether sentence B is the next sentence or whether it is not the next sentence. This process continues, and the model learns from the error rates in each prediction until it fully predicts the accurate sequence of sentences in a text (Devlin et al., 2018).

The version of the BERT model using multilingual text is called mBERT (Devlin et al., 2018). mBERT contains the lexical, linguistic, and grammatical knowledge for 104 different languages. The languages in this model were selected because they contain the largest number of Wikipedia entries. mBERT was trained using MLM and NSP, and it implements a monolingual text stream process in which each language's pre-training process is conducted separately. As a result, the feature space for each language is not shared with any other language in the model. mBERT can be used to overcome the problem of data sparsity. For example, Firoozi and Gierl (in press) used mBERT to score essays written in Persian. Persian is a low-resource language that has proven challenging for traditional automated text analysis methods (Roshanfekr et al., 2017). Firoozi and Gierl (In press) compared the result of the mBERT language model with a word-embedding language model. The mBERT model (Quadratic Weighted Kappa =0.84) significantly outperformed the word embedding model (Quadratic Weighted Kappa = 0.75). The Quadratic Weighted Kappa (QWK) is a statistic that measures the agreement between two sets of ratings or classifications. It is commonly used in the field of inter-rater reliability to assess the agreement between human raters or between a human rater

and an automated system. QWK considers quadratic weights for misclassifications based on how close the ratings are to the correct class using an ordinal scale. In the current study, QWK is the main evaluation metric in AES systems because it can easily be used to compare the performance of our model with the performance of models used in similar studies. QWK varies from 0 (random agreement between raters) to 1 (complete agreement between raters). Typically, values between 0.60 and 0.80 QWK are used as a lower bound estimate for an acceptable reliability outcome using human raters in a high-stakes testing situation (Williamson et al., 2012).

Another BERT model that uses multilingual text is called language-agnostic BERT sentence embedding or LaBSE (Feng et al., 2020). LaBSE is an extension of mBERT where, instead of considering monolingual text streams, bilingual text streams are created by using parallel data—a collection of texts in two or more languages aligned at a sentence or phrase level—in the learning process of the model. This sentence embedding method is called translation language modeling (TLM) (Lample & Conneau, 2019). LaBSE uses TLM and MLM for training by randomly masking words in both the source- and target-language sentences. For example, when Italian and German serve as the source and target languages, respectively, LaBSE can predict a word masked in a sentence written in Italian either by attending to the surrounding words written in Italian or by attending to the parallel surrounding words written in German thereby allowing the model to align the Italian and German text representations.

A critical benefit of using LaBSE is that the model can leverage information from the multilingual context to improve its ability to learn the text (Chi et al., 2020). For example, the model can use the German language context to infer the masked Italian word if the Italian language context is not sufficient to infer the masked Italian words. By training LaBSE on parallel sentences using TLM and MLM, the model learns the lexical, linguistic, and grammatical knowledge for each language and connects the knowledge between the two languages, thereby providing a shared embedding space for both languages in the model. Because LaBSE is pre-trained on 109 languages, many different LaBSE-based language models can be fine-tuned using different numbers and types of languages on downstream tasks. The knowledge transfer between languages is essential when considering low-resource languages. Training a language model like LaBSE on several languages with adequate supervised resources allows building AES systems on low-resource languages using their limited data. LaBSE serves as another example of how language models can be used to facilitate transfer learning (Ranathunga et al., 2023).

GPT (Generative Pre-trained Transformer) is another recent groundbreaking transformer-based language model (Radford et al., 2019) developed by OpenAI. Like the BERT models, the core idea behind training GPT models is the attention mechanism introduced by transformers. GPT models differ from BERT-based models in terms of training methods and the dataset used for training. Unlike BERT, a bidirectional transformer-based architecture, GPT is a unidirectional transformer-based architecture trained on texts from start to end. In addition, GPT models use a different training method than mask language modeling used in BERT. GPT models are autoregressive language models that generate text by predicting the next word in a sequence given the previous words (Black et al., 2022; Brown et al., 2020). This type of training enables GPT models not only to understand but also to generate texts.

GPT models are trained unsupervised on a vast amount of textual data available on the internet. GPT was trained on a much larger corpus than the one used for BERT. For example, GPT-3 (Generative Pre-trained Transformer 3), the third version of the GPT series of language models introduced by OpenAI in 2020, contains 175 billion parameters that enable it to generate coherent and fluent text outputs such as text generation, language translation, and question-answering in a human-like manner. While BERT-based models can mainly be utilized for

transfer learning in scoring students' written tasks, the GPT models' capabilities to generate text make them useful for transferring their knowledge to generate detailed feedback to students (Mayer et al., 2023). GPT-3's remarkable capabilities come with computational resource requirements and limitations. The model size and complexity make it computationally intensive, requiring significant computational power to train and deploy effectively. Additionally, GPT-3 text generation can sometimes exhibit biases in the training data, and it may generate plausible but incorrect or misleading information (Mizumoto & Eguchi, 2023).

GPT is just a member of a larger category of models called Large Language Models (LLMs) (MacNeil et al., 2022). As the name suggests, LLMs are large models trained to contain the structure and knowledge of natural languages. Similar to variations of GPT, these models contain billions of parameters and are trained on massive corpora using self-supervised methods. As a result, these models acquire a deep and rich understanding of their target languages. However, as these models are trained on text with a wide range of topics and structures, they have gained a unique generalizability and multitasking ability (Bubeck et al., 2023). These models can be prompted and interactively trained to perform entirely new tasks. One example of such a task could be AES. Like humans, the model will gradually acquire the ability to do AES by prompting a language model to score an essay and providing constructive feedback. LLMs can utilize their comprehensive knowledge of language, common sense, and communication skills to acquire AES skills without needing to be explicitly trained on AES in a supervised fashion (Mizumoto & Eguchi, 2023). Hence, LLMs can be trained as an AES chatbot that scores essays and provides detailed and personalized feedback for each essay. The chatbot can also be prompted for further feedback and automated improvements through a natural language conversation.

3. AUTOMATED ESSAY SCORING IN TURKISH LANAGUGE

Turkish is a language in the Turkic family of Altaic languages, which over 80 million people speak in Turkey, the Middle East, and Western European countries. Despite being the native language of more than 80 million people, like other low-resource languages, Turkish is also relatively less studied and benefited from the developed NLP tools and resources (Oflazer & Saraçlar, 2018). The Turkish language has certain morphological features, such as multiple derivations of a given word via prefixes and suffixes, making language processing more challenging (Koskenniemi, 1983). For example, the single word "ruhsatlandırılmamak" includes five suffixes. Despite these language challenges, NLP research and tools in the Turkish language are growing thanks to the unsupervised learning algorithms that overcome the problem of data sparsity in NLP tasks, such as speech recognition (e.g., Arslan & Barışçı, 2020) and sentiment analysis (e.g., Gezici & Yanıkoğlu, 2018).

Research on Turkish AES has been the focus of very few studies (Cetin & Ismailova, 2019; Dikli, 2006; Uysal & Doğan, 2021). Cetin and Ismailova (2019) attempted to develop a language tool to automatically evaluate students' essays in Turkish. They used the existing NLP tools for the Turkish language, including Zemberek (Akın & Akın, 2007), to extract mechanical features of the language, such as word count, spelling error, and number of sentences for evaluation of written essays. In another study, Uysal and Doğan (2021) compared different machine learning (ML) algorithms, including support vector machines, logistic regression, multinominal Naive Bayes, long-short term memory (LSTM), and bidirectional long-short term memory (BILSTM) to score open-ended response items in the Turkish language. They also used the existing NLP tools in the Turkish language for text representations. Uysal and Doğan (2021) concluded that the BLSTM model outperformed (QWK=0.77) the other models, including Logistic regression (QWK=0.70), Naïve Bayes (QWK=0.64), support vector machine (QWK=0.69), and LSTM (QWK=0.58) in terms of scoring accuracy.

Despite the popularity of AES models, they have not been studied widely in the Turkish language. One reason is that the NLP tools for feature extraction in low-resource languages such as Turkish are limited (Cetin & Ismailova, 2019). In addition, there are very few, if any, labeled essays available for public research. For example, in the Turkish language, the available few labeled data (e.g., Benchmark Data[†]) are developed for text analysis tasks, such as sentiment analysis (Kavi, 2020), and there are no labeled essays that can be used for AES tasks. Given that the recent large language models such as mBERT and LaBSE are trained in hundreds of languages, including Turkish, the rich knowledge in these LLMs can be transferred to downstream tasks such as AES using even a few training data (Firoozi et al., 2022). The existing challenges in the Turkish AES research, including the limited NLP tools for feature extraction and the insufficient labeled essays available for public research, can be solved by using the large language models, such as mBERT and LaBSE, which were reviewed in the current study. Using transformers, like mBERT, can help decrease the gap in the AES literature between English and low-resource languages. The following steps summarize the process of applying the mBERT model to the Turkish Language using Python.

3.1. Installing Transformers Library

Google Colab (https://colab.research.google.com/) gives free access to writing and executing arbitrary Python code through the browser. It also provides easy-to-use hardware acceleration for deep learning models. First, on the Google Colab page, we install transformers with pip package manager and import the installed packages (Figure 1). The Tensorflow package is a computational graph processor–a fundamental tool for implementing and using deep learning models in Python. The Transformers package by Hugging Face also enables us to employ the latest transformer models and their pre-trained weights.

Figure 1. *Installing libraries.*

```
!pip install tensorflow
!pip install transformers

import tensorflow as tf
from transformers import AutoTokenizer
from transformers import TFAutoModelForSequenceClassification
```

3.2. Loding Turkish Dataset

The code snippet in Figure 2 shows how to write a Python function to read the Turkish AES dataset and return two lists: one containing the texts and one containing the labels. The Turkish AES dataset is a collection of texts in the Turkish language and their corresponding AES scores.

Figure 2. *Loading datasets.*

```
def read_texts(path=ADDR_TURKISH):
    with open(path, "rb") as file:
        dataset_turkish = pickle.load(file)
    texts = [item[0] for item in dataset_turkish])
    labels = [round(float(item[1]) for item in dataset_turkish])

    return texts, labels
```

[†] https://www.kaggle.com/datasets/savasy/ttc4900

3.3. Data Preprocessing

For tokenization, we can use either mBERT Tokenizer or one of the existing Turkish-specific language models, such as BERTurk (https://huggingface.co/dbmdz/bert-base-turkish-cased), using the codes in Figure 3.

Figure 3. *Data preprocessing*.

```
 tokenizer = AutoTokenizer.from_pretrained(
    " bert-base-multilingual-cased"
)
Or
tokenizer = AutoTokenizer.from_pretrained(
    " dbmdz/bert-base-turkish-cased"
)
```

Text tokenization is the process of splitting a text into smaller units, such as words or subwords, that can be mapped to numerical representations. Different methods of text tokenization for transformers are:

Byte-Pair Encoding (BPE): This method uses a statistical algorithm to learn a fixed-size vocabulary of subword units from a large corpus of text. It starts with a set of characters as the initial vocabulary and then iteratively merges the most frequent pair of symbols until the vocabulary reaches the desired size. BPE can handle rare or unknown words by breaking them into smaller subwords. BPE is used by models such as GPT-2 and RoBERTa1.

WordPiece: This method is similar to BPE, but instead of merging the most frequent pair of symbols, it merges the pair that maximizes the likelihood of the data. WordPiece also uses a special symbol to mark the beginning of a word so that it can distinguish between different occurrences of the same subword in different words. WordPiece can also handle rare or unknown words by breaking them into smaller subwords. WordPiece is used by models such as BERT and DistilBERT1.

SentencePiece: This method is a generalization of BPE and WordPiece, which can operate on raw texts without pre-tokenization or pre-segmentation. SentencePiece can learn a vocabulary of subword units from any language and encode texts into sequences of subwords or characters. SentencePiece can also handle rare or unknown words by breaking them into smaller subwords or characters. SentencePiece is used by models such as ALBERT and XLNet1. An example of how a word piece tokenizer—which is used in our sample—might work on a Turkish language sentence is as follows. In the sentence “Türkiye'nin en büyük şehri İstanbul'dur,” means “The largest city of Turkey is Istanbul,” the word piece tokenizer would first split the sentence into words by whitespace as: [Türkiye'nin, en, büyük, şehri, İstanbul'dur.]

3.4. Model Training

BERTurk[‡] is a cased BERT model for the Turkish language that can be used for tokenization. The model is trained on various free-access Turkish corpora, including a filtered and sentence-segmented version of Turkish open parallel corpus (OPUS), OSCAR corpus, and a special local corpus. The final training corpus has a size of 35GB and 44,04,976,662 tokens§. The following codes can be used for tokenization using BERTurk.

To load the BERTurk model using the Hugging Face’s transformers package, we need to use the AutoModel and AutoTokenizer classes from the transformers module. Given the model’s

[‡] https://github.com/stefan-it/turkish-bert
[§] https://huggingface.co/dbmdz/bert-base-turkish-cased

name or path, these classes can automatically load any model from the huggingface model hub. The BERTurk model is available on the model hub under "dbmdz/bert-base-turkish-cased." We can load this model using the codes in Figure 4.

Figure 4. *Importing a pretarined model.*

```
 from transformers import AutoModel, AutoTokenizer


tokenizer = AutoTokenizer.from_pretrained("dbmdz/bert-base-turkish-cased")
model = AutoModel.from_pretrained("dbmdz/bert-base-turkish-cased")
```

The codes in Figure 5 is an example of how to train the BERTurk model using PyTorch. The code defines the loss function, optimizer, and scheduler for the training process. The loss function is a cross-entropy loss, which measures how well the model predicts the correct class for each text. The optimizer is an AdamW optimizer–a stochastic gradient descent method that updates the model parameters based on the gradients of a loss function. The scheduler is a linear schedule with a warmup, which adjusts the learning rate during the training process. The code also defines a training loop, which iterates over the batches of data and labels, feeds them to the model, computes the loss, and updates the model parameters using the optimizer and the scheduler.

Cross-entropy loss is a standard loss function used in machine learning, especially for classification tasks. It measures how well a model predicts the correct class for a given input by comparing the probability distribution output of the model with the true distribution of the classes. The lower the cross-entropy loss, the better the model predicts the correct class. In the context of AES, cross-entropy loss can be used to train a model that assigns scores to essays as an alternative to human grading. AES is a challenging task that requires a model to understand the content, structure, and style of an essay, and to compare it with a predefined rubric or criteria. One way to approach this task is to formulate it as a classification problem, where each possible score is treated as a class. For example, if the scoring scale is from one to fix, then there are six classes to predict.

Figure 5. *Model training.*

```
# Define data and labels
texts = ["Bu bir örnek cümledir.", "Bu başka bir örnek cümledir.", ...] # Your texts here
labels = [0, 1, ...] # Your labels here


# Encode data and labels
inputs = tokenizer(texts, padding=True, truncation=True, return_tensors="pt")
labels = torch.tensor(labels)


# Create data loader
batch_size = 32
data_loader = DataLoader(list(zip(inputs["input_ids"], inputs["attention_mask"], labels)), batch_size=batch_size)


# Define loss function, optimizer, and scheduler
loss_fn = CrossEntropyLoss()
```

```
optimizer = AdamW(model.parameters(), lr=2e-5)
total_steps = len(data_loader) * epochs
scheduler = get_linear_schedule_with_warmup(optimizer,
num_warmup_steps=0, num_training_steps=total_steps)

# Define training loop
epochs = 4
for epoch in range(epochs):
  # Train model on batches of data
  for batch in data_loader:
    # Get batch data and labels
    input_ids, attention_mask, labels = batch

    # Forward pass
    outputs = model(input_ids=input_ids, attention_mask=attention_mask)
    logits = outputs[0]

    # Compute loss
    loss = loss_fn(logits, labels)

    # Backward pass and update parameters
    loss.backward()
    optimizer.step()
    scheduler.step()

    # Reset gradients
    optimizer.zero_grad()
```

3.5. Model Evaluation

The codes in Figure 6 can be implemented to evaluate the BERTurk model on the AES dataset. The code defines the evaluation metrics and writes a function to compute them for a given data loader and model. The code uses accuracy, F1-score, QWK, or Kappa as the evaluation metrics, which measure how well the model predicts the scores of the essays. The code then writes a function that loops over the batches in the data loader, computes the logits (output scores) of the model, gets the predicted labels by taking the argmax of the logits, and calculates the metrics for the predictions and the true labels. The code then runs this function on the test and validation sets and prints the results.

The evaluation metrics used in the context of automated essay scoring are measures of how well the automated system can mimic human raters in grading essays. Each metric captures a different aspect of writing quality and can be used to compare the performance of different models or systems. Here is a brief explanation of why each metric was used:

Figure 6. *Model Evaluation.*

```
# Import the libraries
from sklearn.metrics import accuracy_score, f1_score, cohen_kappa_score

# Define the evaluation metrics
metrics = {"accuracy": accuracy_score, "f1": f1_score, "kappa": cohen_kappa_score}

# Write a function to compute predictions for a given data loader and model
def evaluate(dataloader, model):
  # Set the model to evaluation mode
  model.eval()
  # Initialize empty lists to store the predictions and the true labels
  preds = []
  truths = []
  # Loop over the batches in the data loader
  for batch in dataloader:
    # Get the inputs and labels from the batch
    input_ids = batch["input_ids"]
    attention_mask = batch["attention_mask"]
    labels = batch["labels"]
    # Move them to the device (cpu or gpu)
    device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
    input_ids = input_ids.to(device)
    attention_mask = attention_mask.to(device)
    labels = labels.to(device)
    # Compute the logits (output scores) with no gradient calculation
    with torch.no_grad():
      outputs = model(input_ids, attention_mask)
      logits = outputs.logits
    # Get the predicted labels by taking the argmax of the logits
    pred_labels = torch.argmax(logits, dim=1)
    # Append the predictions and the true labels to the lists
    preds.extend(pred_labels.tolist())
    truths.extend(labels.tolist())
  # Compute the metrics for the predictions and the true labels
  results = {}
  for name, metric in metrics.items():
    results[name] = metric(truths, preds)
  # Return a dictionary of results
  return results

# Run the evaluation function on the test and validation sets and print the results
test_results = evaluate(test_loader, model)
valid_results = evaluate(valid_loader, model)
print("Test results:")
print(test_results)
print("Validation results:")
print(valid_results)
```

Accuracy: Accuracy is the simplest and most intuitive metric. It measures how often the automated system assigns the same score as the human rater. Accuracy is easy to calculate and interpret, but it does not account for the variability or agreement among human raters, nor does it reflect the severity of errors made by the system.

F1-score: The F1-score is the harmonic mean of precision and recall. Precision measures how many of the essays scored by the system are correct, while recall measures how many of the correct essays are scored by the system. F1-score balances both aspects and gives a higher score to precise and recall-oriented systems. F1-score is useful for evaluating systems that assign binary or categorical scores, such as pass/fail or low/medium/high.

QWK or Kappa: QWK or kappa is a measure of agreement between two raters that accounts for the chance agreement. It compares the observed agreement with the expected agreement under random scoring. QWK or kappa ranges from -1 to 1, where 1 means perfect agreement, 0 means no agreement beyond chance, and negative values mean worse than a chance agreement. QWK or Kappa is useful for evaluating systems that assign ordinal or numerical scores, such as 1 to 6 or 0 to 100. It also takes into account the magnitude of disagreement, such that a small difference in scores is less penalized than a large difference.

4. DISCUSSION and CONCLUSION

Pre-trained language models, encompassing both word embedding techniques and transformer-based architectures, present a robust foundation for harnessing extensive knowledge to enhance the efficacy and efficiency of AES models (Singh & Mahmood, 2021). This is especially pertinent in scenarios where data is scarce or challenging to procure. BERT-based language models, in particular, offer a versatile and potent framework for capitalizing on large-scale pre-training to optimize the performance of AES models, even in resource-constrained environments.

The inherent flexibility of BERT-based models extends beyond mere performance enhancement (Devlin et al., 2018). These models facilitate knowledge transfer across languages by undergoing training on a shared set of parameters. Subsequently, this knowledge can be fine-tuned for specific languages or tasks. Our paper succinctly encapsulates key language models that have transformative implications for AES applications in both English and non-English contexts. Furthermore, we expound upon the practical application of mBERT in the Turkish language, displaying its adaptability across linguistic landscapes. As a forward-looking proposition, this research lays the groundwork for future endeavors to implement the methodologies outlined herein. Researchers can utilize the provided codebase to analyze essays written in Turkish, potentially culminating in the development of the inaugural Turkish AES system employing large language models."

Future studies can explore the applicability of alternative transformer-based models, including LaBSE and GPT, to assess their efficacy within the Turkish language. Furthermore, delving into the ramifications of domain-specific fine-tuning on these models' performance in the realm of Turkish essay scoring holds promise for yielding valuable insights.

The scalability of the proposed methodology across diverse languages, coupled with its adaptability to various educational levels and essay genres, opens compelling avenues for subsequent research. Undertaking comparative studies that scrutinize different language models in terms of computational efficiency, interpretability, and bias mitigation could significantly contribute to honing the selection of models tailored for specific AES applications (Yang et al., 2020).

In conclusion, the substantial language models expounded upon in this study serve as a springboard for future AES research across a spectrum of linguistic and educational contexts. Harnessing the capabilities of large language models can empower researchers to actively contribute to the evolution of sophisticated and flexible AES systems, effectively tackling the distinct challenges posed by diverse languages and educational landscapes.

Declaration of Conflicting Interests and Ethics

The authors declare 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 authors.

Contribution of Authors

Tahereh Firoozi: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Okan Bulut**: Methodology, Supervision, and Writing-original draft. **Mark J. Gierl**: Methodology, Supervision, and Writing-original draft.

Orcid

Tahereh Firoozi https://orcid.org/0000-0002-6947-0516
Okan Bulut https://orcid.org/0000-0001-5853-1267
Mark J. Gierl https://orcid.org/0000-0002-2653-1761

REFERENCES

Akın, A.A., & Akın, M.D. (2007). Zemberek, an open source NLP framework for Turkic languages. *Structure*, *10*(2007), 1-5.

Arslan, R.S., & Barişçi, N. (2020). A detailed survey of Turkish automatic speech recognition. *Turkish Journal of Electrical Engineering and Computer Sciences*, *28*(6), 3253-3269.

Bird, S. (2006, July). NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions* (pp. 69-72).

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., ... & Weinbach, S. (2022). Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.

Bouschery, S.G., Blazevic, V., & Piller, F.T. (2023). Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management, 40*(2), 139-153.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Cai, D., He, X., Wang, X., Bao, H., & Han, J. (2009, June). Locality preserving nonnegative matrix factorization. In *Twenty-first International Joint Conference on Artificial Intelligence*.

Cetin, M.A., & Ismailova, R. (2019). Assisting tool for essay grading for Turkish language instructors. *MANAS Journal of Engineering*, *7*(2), 141-146.

Chi, Z., Dong, L., Wei, F., Yang, N., Singhal, S., Wang, W., ... & Zhou, M. (2020). InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, *32*.

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dikli, S. (2006). Automated essay scoring. *Turkish Online Journal of Distance Education*, *7*(1), 49-62.

Firoozi, T., Bulut, O., Epp, C.D., Naeimabadi, A., & Barbosa, D. (2022). The effect of fine-tuned word embedding techniques on the accuracy of automated essay scoring systems using Neural networks. *Journal of Applied Testing Technology*, *23*, 21-29.

Firoozi, T., & Gierl, M.J. (in press). Scoring multilingual essays using transformer-based models. Invited chapter to appear in M. Shermis & J. Wilson (Eds.), *The Routledge International Handbook of Automated Essay Evaluation*. New York: Routledge.

Firoozi, T., Mohammadi, H., & Gierl, M.J. (2023). Using Active Learning Methods to Strategically Select Essays for Automated Scoring. *Educational Measurement: Issues and Practice*, *42*(1), 34-43.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*.

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S., & Köller, O. (2020). Is a long essay always a good essay? The effect of text length on writing assessment. *Frontiers in Psychology*, *11*, 562462.

Gezici, G., & Yanıkoğlu, B. (2018). Sentiment analysis in Turkish. In K. Oflazer & M. Saraçlar (Eds.) *Turkish Natural Language Processing. Theory and Applications of Natural Language Processing* (pp. 255-271). Springer, Cham.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193-202.

Han, T., & Sari, E. (2022). An investigation on the use of automated feedback in Turkish EFL students' writing classes. *Computer Assisted Language Learning*, 1-24.

Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. *Neural Computation*, *9*(8):1735–1780.

Hussein, M.A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*, e208.

Kavi, D. (2020). Turkish Text Classification: From Lexicon Analysis to Bidirectional Transformer. *arXiv preprint arXiv:2104.11642*.

Kenton, J.D.M.W.C., & Toutanova, L.K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of naacL-HLT*, *1*(2).

Koskenniemi K (1983) Two-level morphology: A general computational model for word-form recognition and production. PhD dissertation, University of Helsinki, Helsinki.

Kuyumcu, B., Aksakalli, C., & Delil, S. (2019, June). An automated new approach in fast text classification (fastText) A case study for Turkish text classification without pre-processing. In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval* (pp. 1-4).

Liu, P., Joty, S., & Meng, H. (2015, September). Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods İn Natural Language Processing* (pp. 1433-1443).

MacNeil, S., Tran, A., Mogil, D., Bernstein, S., Ross, E., & Huang, Z. (2022, August). Generating diverse code explanations using the gpt-3 large language model. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 2* (pp. 37-39).

Mayer, C.W., Ludwig, S., & Brandt, S. (2023). Prompt text classifications with transformer models: An exemplary introduction to prompt-based learning with large language models. *Journal of Research on Technology in Education*, *55*(1), 125-141.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050.

Oflazer, K., & Saraçlar, M. (Eds.). (2018). *Turkish natural language processing*. Springer International Publishing.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543).

Ramesh, D., & Sanampudi, S.K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495-2527.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Ranathunga, S., Lee, E.S.A., Prifti Skenduli, M., Shekhar, R., Alam, M., & Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, *55*(11), 1-37.

Rodriguez, P.U., Jafari, A., & Ormerod, C.M. (2019). Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482*.

Roshanfekr, B., Khadivi, S., & Rahmati, M. (2017). Sentiment analysis using deep learning on Persian texts. *2017 Iranian Conference on Electrical Engineering (ICEE*).

Singh, S., & Mahmood, A. (2021). The NLP cookbook: modern recipes for transformer based deep learning architectures. *IEEE Access*, *9*, 68675-68702.

Uysal, I., & Doğan, N. (2021). How Reliable Is It to Automatically Score Open-Ended Items? An Application in the Turkish Language. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(1), 28-53.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Williamson, D.M., Xi, X., & Breyer, F.J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2-13.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A.M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38-45).

Yang, R., Cao, J., Wen, Z., Wu, Y., & He, X. (2020, November). Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1560-1569).

International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 164–173
https://doi.org/10.21449/ijate.1368458
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Implications of current validity frameworks for classroom assessment

Ezgi Mor [1,*], **Rabia Karatoprak Erşen** [1]

[1]Kastamonu University, Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Türkiye

ARTICLE HISTORY

Received: Sep. 29, 2023
Accepted: Dec. 14, 2023

Keywords:
Validity,
Classroom assessment,
Educational assessment,
Argument-based approach,
Educational standards.

Abstract: The argument-based approach is the current framework for validity and validation. One of the criticisms is that understanding and applying this approach to practice are complicated and require abstract thinking. Teachers or school administrators in teaching and learning need support in their validation practice. Due to the abstract structure of validity, the test users and instructors who are not familiar with psychometrics may face problems in gathering validity evidence. Especially in classroom assessment, teachers may deal with understanding the complex methods of validation. In line with this need, the purpose of this study is to help instructors validate their assessment practices by providing a pathway to guide them through their validation processes and to make the validation process more obvious in classroom assessment. For this purpose, a checklist including the validity indicators for classroom assessment is developed. In this development process, Sireci's (2020) 4-step validation which is based on AERA et al. (2014) Standards and Bonner's (2013) study as a framework were followed. The validity indicators were composed by simplifying the AERA's standards and the ones which are relevant to classroom assessment were selected. In addition to the standards, the aforementioned studies were investigated and the validity indicators that may be applicable in classroom assessment were determined.

1. INTRODUCTION

In social sciences, the researchers appeal tests in order to gather information about the people for such a wide range of purposes. Educational and psychological tests are widely used by researchers, employers, and psychologists to make many crucial decisions which are diagnosis, treatment, certification, and evaluation. The consequences of these decisions can be high-stakes in individuals' lives such as enrollment in undergraduate programs, or being licensed to practice their jobs. Hence it is a well-known fact that the tests are valued universally, however, the actual value of the tests is determined by the accuracy level of these decisions. This argument was supported by Sireci and Benitez (2023), who stated that the real value of the tests depends on the quality of the test scores and the provided validity evidence related to the recommended usages of the tests.

In educational and psychological assessment, there is more than one problematic issue that should be handled in a detailed way and one of these issues is the validity of the scores. Validity

*CONTACT: Ezgi Mor ✉ ezgimor@gmail.com Kastamonu University, Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

is one of the concepts frequently considered by almost everyone in education, psychology, or social sciences who has collected data and made inferences based on it. Although it may seem like a rather abstract and technical subject that only those in the field of psychometry can comprehend, the definition of validity has actually undergone radical changes throughout its history in order to make it more unified, observable, and operative.

The concept of validity has been discussed since the early 1900s and is stated as the most vital psychometric quality of test scores (Sireci, 2020). Even though it is explained as the degree to which the test measures the quality it aims to measure, there is not a clear and straightforward definition of validity upon which most of the scholars in the field of educational and psychological measurement agree (e.g., Cizek, 2012; Newton & Shaw, 2014, 2016; Markus, 2016). Validity and validation are defined differently in the primary sources of educational and psychological measurement such as Educational Measurement (Brennan, 2006) and Standards for Educational and Psychological Testing (American Educational Research Association [AERA] et al., 2014). Table 1 presents these definitions.

Table 1. *Definitions of validity and validation.*

	Validity	Validation
Kane (2006, p. 17 in *Educational Measurement*)	the extent to which the evidence supports or refutes the proposed interpretations and uses	evaluating the plausibility of proposed interpretations and uses
Standards for Educational and Psychological Testing (AERA et al., 2014, p. 11)	Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests	accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations

The definitions given in Table 1 are made according to the argument-based approach. Even though it dates back to Cronbach (1988), Kane (1992, 2006, 2013) made the argument-based approach more known and accessible (Sireci, 2020). These definitions refer to not only interpretations but also uses of test scores. According to the argument-based approach, if the argument makes sense and is complete, its inferences are plausible, and the challenges about inferences and assumptions are cleared, then the interpretations/uses (IU) can be considered plausible, in other way, valid.

Kane's argument-based approach first appeared in Kane (1992). Even though it has been around for over 30 years, it has not been widely adopted by professionals in practical settings. Authors such as Newton (2013), Newton and Shaw (2014), and Sireci (2013) criticize the argument-based approach such that understanding and applying it to practice are complicated and require abstract thinking. Furthermore, Moss (2013) and (2016) state that it does not address the assessment needs of teachers or school administrators in teaching and learning and they need support in their validation practice. According to Kane (2013), users are responsible for validation in most cases. However, Moss points out that the information from the test may not have sufficient quality as evidence. Instead, the capacity of how local users use the test data determines the quality of data use. Therefore, validation should be a collaborative practice of test developers and test users.

The validity issues have been accepted as a concern of psychometrics for a long time. Due to the abstract definition and structure of the validity, it may be problematic for instructors who are not interested in psychometry and statistics. The ones who are not familiar with psychological testing or psychometry may be confused while studying the definitions and requirements of validity in AERA et al. (2014) Standards. For this reason, there is a need to develop more concrete ways to analyze the validity of scores, especially in the classroom assessment. As a response to this need, in this study, researchers aim to describe and discuss

the latest validity definitions and develop a checklist including the validity requirements that may be applicable in educational settings. Hence, the purpose of this study is to help the instructors validate their assessment practices by providing a pathway to guide them through their validation processes. This paper starts with a summary of the conceptual evolution of validity and validation. It continues with a part investigating the implications of the validation process in educational settings, especially in classroom assessments based on the research of Bonner (2013) and Sireci (2020). Upon all of the theoretical discussions and analyses, a checklist proposed by the authors was presented followed by the conclusion.

1.1. Validity and Validation

Theoretical discussions about the validity concept may be traced back over a century with Thorndike's (1904) thoughts. His thoughts were accepted as prime for standardized testing in the United States and many European countries (Sireci, 2009). Upon Thorndike's studies, the other prominent development in the concept of validity was observed in the 1940s and 1950s. It was the first time that the researchers reached a consensus about the validity in the Technical Recommendations for Psychological Tests and Diagnostic Techniques: A Preliminary Proposal (APA,1952). In these technical recommendations report for psychological and educational testing, validity was re-conceptualized in different ways and these ways were named as the faces of validity: these are a) content, b) construct, c) predictive, and d) concurrent validity. This report may be accepted as the primary version of professional guidelines on test development, use, and evaluation. After this report, Cronbach and Meehl published a paper in order to discuss the construct validity in 1955 and this research was considered as one of the most influential studies covering the validity issue in a detailed way. Cronbach and Meehl (1955) proposed construct validity as a framework for interpretations about traits that are defined in terms of performance or tasks or behaviors shown by the individuals who have the trait interpreted in terms of lasting characteristics of individuals. The framework was not easy to apply in practical settings but it was influential in setting the term construct instead of trait.

Conceptualizing validity as a unified approach has been tried for a long time. In the 1970s, application-specific practical measurement settings used validity types separately whenever they were appropriate for the interpretations or uses (Kane & Bridgeman, 2021). Messick (1989) provided a unified validity framework centered on construct validity. It was much more comprehensive in defining construct validity than Cronbach and Meehl (1955). However, it was still difficult to apply.

Messick (1980) considers validity as an evaluation and Cronbach (1988) suggests applying the logic of evaluation argument as a framework for validation. Cronbach (1988) connects evaluation with an argumentative approach such that the argument should connect "concepts, evidence, social and personal consequences, and values" (p. 4). Argument-based approach (Cronbach, 1988) provides a framework for validation when there is no formal theory of construct. This approach helps eliminate ambiguity and open-endedness from the validation process by specifying a validity argument. The validity argument provides an evaluation of the proposed score interpretation and use through investigating any evidence related to attempted claims.

Kane (2013) proposes a way to make the reasoning behind claimed score interpretations more explicit and clearer so that the evaluation of that reasoning becomes more manageable. He suggests developing an Interpretation/Use Argument (IUA), where both interpretation and use of scores have equal importance. This is in contrast to his previous (2006) work where the emphasis was only on score interpretations. Kane defines IUA as "a network of inferences and assumptions leading from the test performances to the conclusions to be drawn and to any decisions based on these conclusions" (p. 8). According to this definition, interpretations consist of claims about a unit of analysis, and score uses include decisions about this unit of analysis.

Both interpretation and use have a major role in test development such that they define the purpose of the test. Even though it may seem that interpretations and uses are distinct, they may not be in practice.

The argument-based approach to validity has two steps (Kane, 2013). The first step of validation is to develop the IUA, which helps to understand the required evidence for validation and set the criteria for adequacy of validation. After the IUA has been developed, the IUA is evaluated using a validity argument. A validity argument is the evaluation of the evidence needed to evaluate the inferences and assumptions of the IUA. If the IUA is judged to be clear, coherent, and complete, and its inferences and assumptions are reasonable, then the claimed interpretation or use is valid.

Kane and Bridgeman (2021) stated that there exists an incompatibility between unified and application-specific frameworks since the early conceptualizations of validity. Standards for Educational and Psychological Testing published in 1974 and 1999 could not provide a solution for this incompatibility. 1985 standards necessitate evidence specific to the IU of interest. However, it did not contain much explicit guidance about combining different kinds of evidence. Although 2014 standards are in line with the argument-based approaches, the chapter on validity is written in terms of five kinds of evidence: "evidence based on test content, on response processes, on internal structure, and relations to other variables, as well as evidence for validity and consequences of testing".

One of the criticisms of the argument-based approach is that it does not address validation of local uses (Moss, 2013, 2016) such that the IUA framework mostly focuses on testing programs and intended uses and does not address how actual IU in local contexts can be validated. For instance, the need to validate the consequences of decisions about improving teaching and supporting learning is an example of the local context. In that case, teachers or school administrators are local users, and the actual IU depends on the purpose of these local users. The purpose might be to enable local users to incorporate the information gathered from the test into instructional practices and use the test results to make decisions about classroom activities, which the current validity theory does not support in a simple way.

2. THE IMPLICATION OF THE VALIDATION PROCESS IN THE EDUCATIONAL SETTINGS

Validity has been a primary concept in educational assessment and in line with validity, the test scores and their usage in educational settings are accepted as essential in the whole education process. Education assessment activities are designed and administered to gather information about students' learning processes in order to detect learning deficiencies and determine the students' achievement levels. Teachers are all convinced about these issues, such as deciding the purpose of tests, and the importance of educational assessments, especially in formative however, they have some problems with validity issues. Many teachers expressed concerns about accountability testing with respect to fairness, accessibility, representativeness, and alignment (Welch, 2021). In this point of view, it is clear that most of the teachers need some support in gathering validity proofs for their tests.

As aforementioned, validity has gained so many meanings throughout the history of psychometry and the debates have continued about the additional meanings and implications of it. Welch (2021) stated that there is a gap in understanding validity issues between teachers and measurement experts and in order to bridge the gap, reframing the messages around validity to help teachers understand the theoretical debates in more observable ways. Alignment of the curriculum, relevance, utility of information, comparability, replicability, stability under different modes, and content representativeness in adaptive tests are all areas that are equally important as alignment. One approach may be to relate additional sources of validity to

elements in the peer-review process of the teacher-made test and the scores obtained from it. In response to this, in this part of the study, the researchers aimed to conceptualize the validity in more concrete ways and paraphrased the already mentioned issues of validity in more observable ways, especially in classroom assessment (CA). While doing this, the resources stated in the first part of the study were used, and especially the AERA standards were benefited mainly. In addition to the APA standards on validity, Bonner's (2013) work on validity in CA was investigated in detail.

2.1. Validity in Classroom Assessment

Bonner (2013) asserted that CA differs from other educational assessments in a radical way in terms of purposes, hence validity may be a secondary purpose for CA. The researcher also stated that in CA, validity or appropriateness of inferences about test scores should be the real concern and it is recommended that teachers and researchers may use the validity analyzing methods to judge the propriety of the inferences.

Bonner (2013) proposed five critical principles that may be used in CA and if these criteria are taken into account, the researcher claims that the sensitivity to individual learners and learning outcomes may be reflected in the assessment process. Also, these principles are equally relevant to validity claims of the researchers and both types of data; qualitative and quantitative. These criteria are listed below:

1. Assessment should be aligned with instruction: It is stated that the curricular standards are not enough for achievement tests in CA. The tests should be aligned based on the tasks used in instruction. Nitko (1989) also supported this idea long before Bonnes (2013) study by defining the appropriate uses of tests that are linked with or integrated with instructional materials and procedures. Bonnes (2013) improves this claim by stating that if the CA is aligned with the instruction poorly, CA may have negative impacts on students' attitudes, motivation, and classroom environment. It is suggested to analyze test content represented on a test by comparing the instruction time or emphases on lesson plans.

2. Bias should be minimal at all phases of the assessment process: This criterion is so crucial, especially for the multicultural classroom environment. Students are open to many diverse factors' effects on the testing process. Some items may be in favor of fluent readers in paper-and-pencil tests, glib writers in essay formats, and personality attributes and performance assessments. Also, the teachers may be affected by biases when scoring the items. In order to minimize the influences of bias in CA, which also increases the validity of CA, tests and tasks can be analyzed by subject-matter experts, a group of teachers, or reviewed and debriefed assessments with a small group of students. Methods to reduce scoring bias, use of rubrics, co-scoring, and multiple-raters for samples of student work may be preferred.

3. Assessment processes should elicit relevant substantive processes: Thinking processes and task-relevant behaviors that are consistent with cognitive perspectives on assessment should be included in CA. Using cognitive processes in the tests may provide better diagnostic information about students' learning levels. Also, these cognitive processes should be included not only in tests but also in scoring phases by using rubrics.

4. Effects of assessment-based interpretations should be evaluated: The results and decisions based on test scores should be justified by strong logical arguments or evidence. Both cognitive and affective consequences of the tests should be analyzed. Especially for formative assessments, teachers should attempt to provide opportunities for students to be reassessed if the results of tests are ineffective or inappropriate.

5. Validation should include evidence from multiple stakeholders: Teachers should know and accept that the validity of their assessment-based decisions, but these decisions may be questioned by the other stakeholders. However it is a fact that there is no requirement of the

getting the approval of all the stakeholders' about the CA decisions. Kane (2006) emphasized the importance of the other stakeholders, who are not in the development processes of the tests, including the consequences of the tests, without this inclusion, the assumption that our assessment-based decisions are all valid. Teachers, who are in the development process of CA, are primarily responsible for evaluating their assessment processes and the assessment-based consequences. Hence the stakeholders may be colleagues, mentors, or professional test developers. As a principle, responsibility for assessment validation should be dependent on the judgment of a single individual.

These five criteria emphasize the importance of validity in CA and teachers are able to apply most of the stated procedures in order to validate the test scores. The other research that focuses on the validation process in a more applicable way is Sireci's (2020) work. In the following part of the study, the research is presented and Sireci's (2020) stepwise perspectives on the validation process are analyzed.

2.2. Sireci's Validation Steps

In the previous part, the five criteria proposed by Bonner (2013) were explained in detail, and it is a fact that these criteria do not differ radically from the AERA et al. (2014) Standards. Actually, most of the validity studies are based on these standards, and one of the most prominent and current studies investigating the validity in line with the AERA et al. (2014) Standards is Sireci's (2020) work. In this research, the researcher investigated the history of the validity concept and updated his previous Sireci (2013) study for the validation process. In Sireci (2013), the researcher proposed a three-step validation process based on AREA et al. (2014) Standards. These steps involved 1) clear articulation of testing purposes, 2) consideration of potential test misuse, and 3) crossing test purposes and potential misuses with the Standards' five sources of validity evidence. In the updated study, Sireci (2020) added one more step and it is 4) prioritizing the validity of studies to be conducted. In this part, these steps were explained concisely and the validity investigation ways that may be adapted in CA were emphasized especially.

Step 1. Articulating the Purposes of the Test: The process of validation includes gathering and analyzing evidence in order to defend the purpose of test usage. In line with the AERA et al. (2014) Standards of validation, Sireci (2020) also emphasized that the validation process begins with the explicit statement of the proposed interpretations of the test scores and of course, this purpose should be supported by a rationale. The important issue is that the intended purposes should be defined in an explicit and concise way and most of the time, the purposes are composed in a general, unclear, and complex way.

Step 2. Identifying Potential Negative Consequences of Test Use: As Messick (1989) stated, it is not enough to determine the intended test usage. It is also crucial to define the potential negative effects of the testing programs. Sireci (2020) suggested criticizing testing programs' adverse effects at the public level. For the large-scale assessment test, it may be stated that it has the potential to influence the curriculum negatively. These potential negative effects should be investigated at test level.

Step 3: Crossing test purposes and potential misuses with the Standards' five sources of validity evidence: In this step, the sources of validity evidence defined in the standards were included. These sources are test content, response processes, internal structure, relations with other variables, and testing consequences. The sources are explained in detail in the Standards, and Sireci (2020) exemplified their usages of them in the validation process with the Massachusetts Adult Proficiency Tests (MAPT) by using the technical manual of this test. Upon analyzing the questions; the ones that may be related with the CA were found and given below by adapting the CA settings:

1. Does the test actually measure students' achievement/ability/ skill /knowledge in the related course?
2. Does it measure these knowledge and skills as they are defined in the curriculum framework?
3. Are the test scores useful for evaluating students' progress toward meeting educational goals?
4. Are the test results useful for evaluating the related program/curriculum of the course?
5. What are the effects of the test on instruction in the education process?

The questions stated above do not stem from only the explicit testing purposes of test use but, some of them especially the last one emanates from implied test purposes, too. However, in the CA, nearly all of the stated questions should be investigated by the teachers who developed a test.

Step 4: Prioritizing the Validity Studies to be conducted: It is a well-known fact that all validity evidence suggested by the standards and/or the related research, are not possible to be gathered, hence some prioritization is needed in order to use time and resources efficiently. This prioritization should be applied based on the purpose of the test and sufficient validity evidence should be gathered as parallel with the test's purpose.

This four-step approach serves as the investigation of validity in an argument-based approach and within this approach, Sireci (2020) emphasized that the limitation of this approach is that it requires responsible test developers and evaluators to clearly articulate testing purposes and the intended information. The other drawback of this approach is stated that applying this approach requires prioritization and it may be problematic to select the type of validity evidence to be gathered. Of course, gathering all types of validity evidence and answering all the research questions for validation is not feasible and that's why prioritizing research questions is needed (Sireci, 2020).

Despite the hottest debates on the approaches and definitions of validity, it is clear that there are still several open-ended and questionable points of the validity investigations for the researchers. Actually, the validity issues may be analyzed in a more direct and easier way for educational settings because the tests used in classroom assessment are developed mainly for determining learning levels and monitoring students' progress. Hence the purposes of teacher-made tests are more obvious and the validity evidence may be gathered easily. In order to make the validation process more trackable and objective, the validity indicators that may be efficient in CA were determined and prepared as items that are open to be questioned by teachers or instructors who developed the tests. These indicators are presented below:

2.3. Validity Indicators in CA

In this part, the determined validity indicators are given. While composing this checklist, the researchers studied collaboratively and the draft of the checklist was analyzed by two different measurement specialists, who had doctorate degrees in measurement and assessment. Based on the experts' views, the indicators were prepared as a form of the checklist format which is composed of 17 items with three grading categories, satisfied, not-satisfied and not applicable. This checklist is presented in Table 2 below.

The validity indicators were prepared to cover the whole validation process. Hence, the checklist was composed by adopting an inclusive approach in which the whole validation process was considered. If the items are investigated in detail, the validation process can be observed. The checklist starts with the definition of the main purpose of the test, which is the first step of the development process of any test. The second and third items are aligned with the first one, it is stated that the possible usages of the test should be described in a detailed way. This is specifically important when the current validity frameworks (e.g., Kane, 2013; Sireci, 2013) include both score interpretations and score uses in the validation process. The third item is closely related to the first indicator in which teachers/ instructors are expected to

relate all the test items by considering the main purpose of the test. In the fourth item, characteristics of the test takers are emphasized and the test-takers should be defined in a detailed way. The fifth and sixth items are related to the scores' meaning and these items emphasize the usage of the scores. These items are so essential that the total score of the test is expected to reflect the level of the measured trait, which depends on the items' scoring. The next two items, seventh and eighth, are related to organizing the administration process of the test. Then in the next items, the content of the items and the relationships among the items are also considered and the determination of item formats is included in the validation process. The item formats should be selected as parallel with curriculum and teaching activities. The scoring criteria and weighting of the items are also included in the validation process. Lastly, the reliability evidence was emphasized in the context of the validation process. In brief, with these indicators, we exerted to cover all the validation steps which were determined according to the primary sources such as AERA et al. (2014) standards, Kane (2006, 2013), Bonnes (2013) and Sireci (2020). These indicators are suggested to be essential for the tests used in CA made by teachers/instructors

Table 2. *Validity indicators checklist.*

Validity Indicators	Satisfied	Not Satisfied	Not Applicable
1. The main purpose of the test is defined.			
2. The proposed test uses are stated in a detailed way.			
3. The test is designed in order to measure students' features; such as achievement/ability/skill/knowledge.			
4. The group of students for which the test is intended is specified.			
5. Test scores are composed to provide useful information for evaluating students' progress toward meeting educational goals.			
6. Test scores are composed to provide useful information for evaluating the related program of the course.			
7. Test administration procedures are determined before the test administration.			
8. The procedures followed in generating test content are justified.			
9. Both the item formats and the content of the items are aligned with the curriculum.			
10. In addition to the included content domains, the areas of the content domain that are not included are indicated.			
11.The test scoring procedures are described in detail.			
12. If it is claimed that the test is unidimensional, such a claim is justified with statistical analysis.			
13.The relationships among the items are investigated using item scores.			
14. Reliability evidence for each reported score is provided.			
15. If a test provides more than one score, the distinctiveness of the separate scores is justified.			
16. If a test provides a composite test score, the basis and how the test scores are combined are justified.			
17. If a differential weighting is proposed by test developers/teachers, the rationale behind the scoring is specified.			

3. CONCLUSION

The validity chapter written by Kane in the current edition of *Educational Measurement* (Brennan, 2006) and *Standards for Educational and Psychological Testing* (AERA et al, 2014) adopt the argument-based approach for validity and validation. According to Kane (2013), it is flexible in accommodating various applications such as achievement testing or experimental designs where causal inferences are made. As long as the claims made according to test scores are plausible and representative of the test scores, and justified empirically, then IUA (i.e., interpretation/use argument) is valid. However, Sireci (2013) argues that the development of interpretive argument, especially scoring, generalization, and extrapolation inferences can be complex and overwhelming, which might discourage practitioners from using the IUA framework. Another criticism is the lack of support for the professionals working in teaching and learning (Moss, 2013; 2016). Sireci (2020) in which 4-step validation using AERA et al. (2014) Standards as a framework for validation practices is proposed can be a practical guidance towards these criticisms.

In the second part of this study, upon evaluating and analyzing the primary studies in this field, the implications of the validation process in educational settings, especially in CA were determined. By investigating the argument-based approach (Kane, 2006; 2013) and Sireci's (2020)'s ideas and suggestions on the validation process, we proposed a checklist in which the essential validation indicators are included. While preparing these items, the clarity and simplicity of the statements were essentially paid attention. Due to the complexity of the Standards, teachers/ instructors may face some problems in understanding and applying these standards in their tests and test scores. Hence, we aimed to develop a short, brief instrument by prioritizing the CA applications and needs. Hence, we aim that the checklist may be used by a wide range of researchers who may be unfamiliar with psychometric issues in depth. With this checklist, teachers or instructors are able to evaluate their test scores by using this checklist, and in order to obtain more valid scores from the tests, they may evaluate their test items, testing conditions, scoring, and the process of test development in terms of these indicators. These indicators are stated as a three-point grading format in which the teachers/instructors may select the appropriate option for their tests and test scores. All items are designed as applicable for all types of tests that may be administered in CA.

Declaration of Conflicting Interests and Ethics

The authors declare 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 authors.

Authorship Contribution Statement

Ezgi Mor: Investigation, Resources, and Writing-original draft. **Rabia Karatoprak Erşen:** Methodology, and Writing-original draft.

Orcid

Ezgi Mor https://orcid.org/0000-0003-0250-327X
Rabia Karatoprak Erşen https://orcid.org/0000-0001-8617-1908

REFERENCES

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

Bonner, S.M. (2013). Validity in classroom assessment: Purposes, properties, and principles. In J.H. McMillan (Ed.), *SAGE handbook of research on classroom assessment*, (pp. 87-106). SAGE.

Cizek, G.J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*(1), 31. https://doi.org/10.1037/a0026975

Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum Associates, Inc.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 174-203.

Kane, M.T. (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.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M., & Bridgeman, B. (2021). The evolution of the concept of validity. In B.E. Clauser & M.B. Bunch (Eds.), *The history of educational measurement* (pp. 181-205). Routledge.

Markus, K.A. (2016). Alternative vocabularies in the test validity literature. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 252-267. https://doi.org/10.1080/0969594X.2015.1060191

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012–1027. https://doi.org/10.1037/0003-066X.35.11.1012

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Moss, P.A. (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement*, *50*(1), 91-98. https://doi.org/10.1111/jedm.12003

Moss, P. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice, 23*(2), 236-251. https://doi.org/10.1080/0969594X.2015.1072085

Newton, P.E. (2013). Two kinds of argument?. *Journal of Educational Measurement*, *50*(1), 105–109. https://doi.org/10.1111/jedm.12004

Newton, P., & Shaw, S. (2014). The deconstruction of validity: 2000–2012. In *Validity in educational and psychological assessment* (pp. 135-182). Sage.

Newton, P., & Shaw, S. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice, 23*(2), 178-197. https://doi.org/10.1080/0969594X.2015.1037241

Sireci, S.G., & Benítez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema 35*(3) 217-226. https://doi.org/10.7334/psicothema2022.477

Sireci, S.G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, *50*(1), 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://doi.org/10.59863/CKHH8837

Welch, C.J. (2021). Rethinking measurement 101: Lessons learned from teachers. *Educational Measurement: Issues and Practice*, *40*(4), 13-17. https://doi.org/10.1111/emip.12479

International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 174–193
https://doi.org/10.21449/ijate.1343248
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Examination of response time effort in TIMSS 2019: Comparison of Singapore and Türkiye

Esin Yilmaz Kogar [1*], **Sumeyra Soysal** [2]

[1]Niğde Ömer Halisdemir University, Faculty of Education, Department of Educational Sciences, Division of Educational Measurement and Evaluation, Niğde, Türkiye
[2]Necmettin Erbakan University, Ahmet Keleşoğlu Faculty of Education, Department of Educational Sciences, Division of Educational Measurement and Evaluation, Konya, Türkiye

ARTICLE HISTORY
Received: Aug. 15, 2023
Accepted: Dec. 12, 2023

Keywords:
Response time effort,
Rapid guessing behavior,
Solution behavior,
Cognitive domain,
Content domain.

Abstract: In this paper, it is aimed to evaluate different aspects of students' response time to items in the mathematics test and their test effort as an indicator of test motivation with the help of some variables at the item and student levels. The data consists of 4th-grade Singapore and Turkish students participating in the TIMSS 2019. Response time was examined in terms of item difficulties, content and cognitive domains of the items in the mathematics test self-efficacy for computer use, home resources for learning, confident in mathematics, like learning mathematics, and gender variables at the student level. In the study, it was determined that all variables considered at the item level affected the response time of the students in both countries. It was concluded that the amount of variance explained by the student-level variables in the response time varied for each the country. Another finding of the study showed that the cognitive level of the items positively related to the mean response time. Both Turkish and Singaporean students took longer to respond to data domain items compared to number and measurement and geometry domain items. Additionally, based on the criterion that the response time effort index was less than .8, rapid-guessing behavior, and therefore low motivation, was observed below 1% for both samples. Besides, we observed that Turkish and Singaporean students were likely to have rapid guessing behavior when an item in the reasoning domain became increasingly difficult. A similar result was identified in the data content domain, especially for Turkish graders.

1. INTRODUCTION

Today's conditions have brought about the necessity of digitalisation in many areas of human life such as trade, health and education. Especially, in the field of education, both courses and exams have started to be conducted with online or computer-based applications, and these applications have become more common and important in recent years. Computer-based applications have also been included in the cycle of large-scale international assessments such as the Trends in International Mathematics and Science Study (TIMSS), the Programme for International Student Assessment (PISA). In 2019, eTIMSS was added to TIMSS as a computer-based "eAssessment system", measuring the same mathematics and science

*CONTACT: Esin YILMAZ KOGAR esinyilmazz@gmail.com Niğde Ömer Halisdemir University, Faculty of Education University, Faculty of Education, Department of Educational Sciences, Niğde, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

constructs using the same assessment items as possible with "paperTIMSS", which is in the paper-and-pencil format as in previous TIMSS cycles (Mullis et al., 2016). Besides, eTIMSS provides more detailed information about students and allows different assessments to be made.

Computer-based testing applications, such as eTIMSS, allow for the collection of a chronology of test takers' interactions with test items throughout the assessment process (Organisation for Economic Co-operation and Development [OECD], 2015). Moreover, such applications make it possible to obtain measures of response time per item, which is difficult to measure in pen-and-paper assessments. The term response time (RT) refers to the time it takes test takers to respond (react) to a particular item (stimulus) in the test (Lee & Chen, 2011). This enables the analysis of test-takers' efforts to take the test through objective records of their actions rather than relying on self-reported assessments of their behaviour (Lee & Jia, 2014; Wise & Kong, 2005). Item response time, which is more easily obtainable through computer adaptive testing, helps understand which factors affect how quickly an examinee answers an item, and thus, helps test developers estimate the time required to answer the total of the test (Bergstrom et al., 1994). The use of process data from large-scale tests, such as eTIMSS, PISA, to determine test effort or grader behaviour has great potential for educational assessment.

The common purpose of large-scale international assessments is to evaluate education systems worldwide by testing and analyzing the abilities and understanding of students of different ages in participating countries/economies. Since the scores on such tests, so-called low-stakes tests, do not indicate any personal conclusions about the test taker's performance, individuals may be reluctant to demonstrate the full range of their knowledge, skills or attitudes. Therefore, since it is unclear whether test takers are motivated enough, test scores may not represent their true ability level and may not serve as a valid measure of their abilities. In this context, many researchers have investigated the function of test-taking motivation during low-stakes assessments of their performance (e.g., Barry & Finney, 2009; Eklöf, 2007; Wise & Kong, 2005; Wise & DeMars, 2005). For this purpose, Wise and Kong (2015) established a relationship between motivation to take the test and response time.

Unlike the studies that relied on examinee self-reports to measure test-taking effort, Wise and Kong (2005) developed a measure, called response time effort (RTE), using item response times, which represents a direct observation of the test taker's behavior and whose collection is unobtrusive and nonreactive (examinees with computer-based test will typically be unaware). Baumert and Demmrich (2001, p.441) defined this term as the following: “test-taking effort as a student’s engagement and expenditure of energy toward the goal of attaining the highest possible score on the test.” Test-taking motivation is also identified as “the willingness to engage in working on test items and to invest effort and persistence in this undertaking”; in short, it is the motivation of the individual to achieve a high-performance level in a test (Eklöf, 2010). Based on the relationship between these two phenomena, Wise and Kong (2005) and Wise (2017) interpreted the RTE as an indicator of test motivation, which contains two types of behaviour. The first of these behaviours is characterized by active engagement in seeking the correct answers to test items, known as solution behavior (SB), while the second is marked by quick responses in a mostly random manner, referred to as rapid-guessing behaviour. Accordingly, the researchers assume that less-motivated examinees are likely to exhibit rapid-guessing behavior, while high-motivated examinees are likely to display SB.

Exams such as the Scholastic Aptitude Test and the American College Test conducted in the USA are high-stake tests where students try getting admission from a college or university, and their performance in these exams directly concerns them (Sundre & Kitsantas, 2004). In contrast, international large-scale assessments classified as low-stake tests (PISA, TIMMS, NAEP, PIRLS, etc.) provide national-level reports without individual results for students, teachers, or parents. Therefore, students' motivation to take the exam may emerge as a problem.

In the literature, the variability of test-taking effort and motivation of the examinees during a low-stakes assessment was examined in various aspects and conditions. Some researchers have focused on the relationship between test-taking effort, test motivation and test performance (e.g., Cole et al., 2008; Lundgren & Eklöf, 2020; Slim et al., 2020; Wise & DeMars, 2005), some researchers have examined predictor variables of examinee's RT (e.g., Baumert & Demmrich, 2001; Gershon et al., 1993; Lundgren & Eklöf, 2020; Wolgast et al., 2020) or properties of the test affected by the RT (e.g., Fan et al., 2012; Wang & Hanson, 2005; Weirich et al., 2017). The number of studies that test effort based on item-response time is more than the ones that are based on self-report measures. The majority of studies employing RT indices have consistently found that a significant percentage of examinees, ranging from 74% to 99%, demonstrated response time values greater than .90. However, it has been noted that the behavior of 1-23% of examinees raised concerns, as they displayed rapid-guessing behaviour on more than 10% of the test items (e.g., Setzer et al., 2013; Swerzewski et al., 2011; Wise & DeMars, 2005; Wise & Kong, 2005). These findings demonstrate that when RT indices were utilized as a measure of test effort, the majority of examinees consistently displayed diligent efforts in answering the items, with minimal within-examinee variation in their levels of effort throughout the test. Wise and DeMars (2005) raised concerns about the potential for examinees to provide inaccurate or insensible responses on self-report scales, perhaps, which may have led to the limited use of self-report methods to examine test-taking efforts in a few studies (Barry et al., 2010; Myers & Finney, 2021; Wolgast et al., 2020).

In the world, which has already been increasingly digitilized for the last few decades, the place and importance of computerized and online learning environments in education has gradually increased with the significant impact of the COVID-19 pandemic. Besides, the effect of digitalization could be seen in international large-scale assessments such as the TIMSS. In the last few cycles of these assessments, a paper-pencil format or a computer-based "e" version has been presented to selected countries. However, over 50% of the 64 nations involved in the TIMSS 2019 opted to conduct the "e" version of the assessments, while the remaining countries followed the traditional approach of administering the TIMSS using pen and paper, as done in previous cycles (Martin et al., 2020).

Although the most basic variable that may have an effect on students' RT is the ability level of the student, different variables may also affect RT. For example, studies in the literature show that RT is also related to different item-level variables such as item difficulty level, content area, and conginitve domain (Bridgeman, & Cline, 2000; Goldhammer et al., 2014; Hess et al., 2013; İlgün-Dibek, 2020; Lee & Jia, 2014; Wang, 2017; Yalçın, 2022; Zenisky & Baldwin, 2006). Besides, student-level variables may also be related to RT. Among such variables, there are studies addressing gender (Hess et al., 2013; İlgün-Dibek, 2020; Setzer et al., 2013) and self-confidence (Yalçın, 2022) in terms of RT. Cooper (2006) and Zhang et al. (2016) reported that test outcomes were affected by students' comfort and self-confidence levels in using computers and tablets. Considering this situation, we wanted to examine the effect of computer self-efficiency computer on response time in our study. In addition, in TIMSS 2019, countries, not students, decided which version of the paper TIMSS or eTIMSS would be implemented in countries. Given that not all individuals have the same opportunities, students who are not familiar with the use of such digital devices may experience difficulties in computer-based assessments (Bennet et al., 2008; Chen et al., 2014; Pommerich, 2004). Since this familiarity is obviously related to home resources, the home resources variable was also considered in the study. Although studies on RT analyses are available in the literature, there are fewer studies on RT analyses for country comparisons (see, İlgün-Dibek, 2020; Rios & Guo, 2020; Michaelides et al., 2020). Therefore, in the current study, we aimed to gather evidence on possible differences in response time efforts between countries, which is intended to increase the validity of cross-country comparisons. Thus, we believe that this study will contribute to

the literature on cross-country comparisons of response time effort. We have summarised the aim and sub-problems of our research in detail below.

1.1. Purpose of the Study

The aim of the current study is to evaluate different aspects of students' RT and test-taking motivation using some item- and student-level variables, based on data from the 2019 TIMSS 4th grade samples from Türkiye and Singapore. In this context, the following subquestions, which the study is intended to answer, are presented as follows:

1) Does the mean response time of students in the mathematics test, each for Türkiye and Singapore samples, significantly differ according to content domain, cognitive domain, and item difficulty to which the items belong?
2) Is the mean response time of students in the mathematics test, each for Türkiye and Singapore samples, significantly predicted by self-efficacy for computer use, home resources for learning, like learning mathematics and gender?
3) How is the response time effort of students in the mathematics test, each for Türkiye and Singapore samples?

2. METHOD

2.1. Datasets

The data of the study consists of 4th-grade Singapore and Türkiye students participating in the TIMSS 2019, which can be downloaded from the International Association for the Evaluation of Educational Achievement (IEA) website. The reason why we included Türkiye and Singapore in this research is that we wanted to compare Singapore, which had the highest performance in mathematics with 625 points, with our country which ranked 23rd with 523 points. 5986 students (47.9% girl) participated in Singapore and 4028 students (52.1% girl) participated within Türkiye. The total number of items analyzed was 159. For 27 derived items where students were asked to give more than one answer or a multi-part answer, the response time (total time on screen as seconds) was divided by the number of items contained in the derived item. Similarly, item difficulty statistic for a derived item was rearranged to represent the mean difficulty of the items it contained.

2.2. Variables in Interest

2.2.1. *Item-level variables*

2.2.1.1. Content Domain (CnD). One of the dimensions, which each of the paper TIMSS and eTIMSS assessment frameworks is organized around and specifies the subject matter to be assessed. In the 4th-grades, a mathematics test consists of three content domains, which are apportioned as follows: number (50%), measurement and geometry (30%) and data (20%) (Martin et al., 2020). In this study, the items were coded as 1 = numbers, 2 = measurement and geometry, 3 = data through the analysis.

2.2.1.2. Cognitive Domain (CD). Paper TIMSS and e-TIMSS assessment frameworks are each structured around a dimension that outlines the specific cognitive processes to be assessed. For all grades, a mathematics test consists of three cognitive domains, which is apportioned as follows: knowing (40%), applying (40%) and reasoning (20%) (Martin et al., 2020). In this study, the items were coded as 1 = knowing, 2 = applying, 3 = reasoning through the analysis.

2.2.1.3. Item Difficulty (*p*). It was calculated by dividing the number of test takers who answered correctly by the total number of test takers. Item percent correct statistics was used from the TIMSS 2019 International Database. Although there are different classifications for item difficulty index, the three-category classification as referred by Crocker and Algina (1986,

p.324) was used in this study to avoid too many categories: hard (0 to .39), moderate (.40 to .60) and easy (.61 to 1.00). For item-level analysis, the distribution of items across the three cognitive domains and the three content domains by item difficulty is summarized in Table 1.

Table 1. *Descriptive statistics for items.*

Country	Cognitive Domain	Content Domain	Item Difficulty			Total
			Hard	Moderate	Easy	
Singapore	Knowing	Numbers	-	-	31	31
		Measurement and Geometry	-	3	15	18
		Data	-	-	8	8
		Total	-	3	54	57
	Applying	Numbers	1	2	34	37
		Measurement and Geometry	-	3	14	17
		Data	-	1	12	13
		Total	1	6	60	67
	Reasoning	Numbers	1	8	4	13
		Measurement and Geometry	3	3	8	14
		Data	1	1	6	8
		Total	5	12	18	35
Türkiye	Knowing	Numbers	3	9	19	31
		Measurement and Geometry	4	7	7	18
		Data	1	3	4	8
		Total	8	19	30	57
	Applying	Numbers	6	21	10	37
		Measurement and Geometry	3	11	3	17
		Data	3	6	4	13
		Total	12	38	17	67
	Reasoning	Numbers	9	3	1	13
		Measurement and Geometry	5	7	2	14
		Data	2	2	4	8
		Total	16	12	7	35

As shown in Table 1, each cognitive domain and content domain contained a considerable number of items. The number of items for Numbers, Measurement and Geometry, and Data content domains is 81, 49 and 29, respectively. The number of items for knowing, applying and reasoning cognitive domains is also 57, 67 and 35, respectively. For Singapore sample, while the percentage of correct answers to the items in the knowing and applying domains by the students is quite high, the items within the reasoning domain are a substantial amount of medium and easy difficulty level. Additionally, it can be said that Singaporean students do not have difficulty in the data domain, but they have some difficulties in the measurement and geometry domain. In the Türkiye sample, the items within knowing were mostly on the easy difficulty level, while the items within applying and reasoning were classified as medium and high. It was observed that Turkish students had more difficulties as the cognitive domain level of the contents increased.

2.2.2. *Student-level variables*

2.2.2.1. Gender. This variable was coded as 1 = Girl and 2 = Boy throughout the analysis.

2.2.2.2. Students Like Learning Mathematics (SLM). The scale has nine items with a 4-point response key ranging from agree a lot to disagree a lot, which covers students' attitudes toward mathematics and studying mathematics. The total scale score is divided into three categories: very much like (score at or above 10.2), somewhat like (between 10.2–8.4) and do not like (at or below 8.4). The percentages of students in Singapore to this variable are as follows: 36.9% very much like mathematics learning, 40.0% somewhat like learning mathematics, and 22.9% do not like learning mathematics. For Türkiye, the classification is as follows: 64.8% very much like learning mathematics, 25.4% somewhat like learning mathematics, and 9.2% do not like learning mathematics.

2.2.2.3. Students Confident in Mathematics (SCM). This scale measures how confident students feel about their ability in mathematics, in terms of their level of agreement with nine statements with a 4-point response key ranging from agree a lot to disagree a lot. The total scale score is divided into three categories: very confident (score at or above 10.7), somewhat confident (between 10.7–8.5), and not confident (at or below 8.5). The percentages of students in Singapore by this variable are as follows: 20.7% very confident in mathematics, 42.0% somewhat confident in mathematics, and 37.1% not confident in mathematics. For Türkiye, the classification is as follows: 33.2% very confident in mathematics, 41.3% somewhat confident in mathematics, and 23.4% not confident in mathematics.

2.2.2.4. Home Resources for Learning (HRL). This measurement scale combines data gathered from fourth-grade students and their parents. The students supplied details regarding the number of books and other study supports in their households, while the parents provided information concerning the number of children's books, the educational levels of the parents, and the occupational status of the parents. The total scale score is divided into three categories: many resources (score at or above 11.8), some resources (between 11.8–7.4) and few resources (at or below 7.4). High scores indicate that the student has more home resources. The percentages of students in Singapore by this variable are as follows: 28.3% many resources, 66.6% some resources, and 1.7% few resources. For Türkiye, the classification is as follows: 4.6% many resources, 64.1% some resources, and 24.3% few resources.

2.2.2.5. Self-Efficacy for Computer Use (SEC). We could not find any detailed information for this variable in the TIMSS 2019 technical report. As data used in this paper in Singapore, 50.0% of the students are in the high self-efficacy, 39.3% of the students are in medium self-efficacy, 2.4% of the students are in low self-efficacy category. In Türkiye, 63.8% of the students are in the high self-efficacy, 33.0% of the students are in medium self-efficacy, 2.6% of the students are in the low self-efficacy category.

2.2.2.6. Response Time Effort (RTE). As explained earlier, the test-taking effort is assessed by analyzing item response times, focusing on two distinct behaviors known as *solution behavior (SB)* and *rapid-guessing behavior*. SB refers to cases where examinees put effort into answering an item thoughtfully. On the other hand, examinees who quickly respond without sufficient time for reading and full consideration of an item exhibit rapid-guessing behavior (Wise & Kong, 2005). Thus, SB is considered effortful response strategies, while rapid guesses are seen as non-effortful strategies. The 10% normative threshold (NT10) methodology proposed by Wise and Ma (2012) was used to determine whether each student showed SB or rapid-guessing behaviour over the answering time. As part of this approach, the initial step involved computing the average response time for each item, and then 10% of this value was used as a threshold. However, according to the recommendation made by Setzer et al. (2013),

it is advised to employ a maximum threshold of 10 seconds when utilizing this methodology. Therefore, we established the maximum threshold at 10 seconds. After one threshold (T_i) was determined for each item, the following steps were followed: For item i, there is a threshold, T_i, that represents the response time boundary between rapid-guessing behavior and solution behavior. Given an examinee *j*'s response time, RT_{ij}, to item *i*, a dichotomous index of item solution behavior, SB_{ij}, is computed as in Equation 1 (Wise & Kong, 2005, pp.167-168).

$$SB_{ij} = \begin{cases} 1 \; if \; RT_{ij} \geq T_i, \\ 0 \quad otherwise, \end{cases} \tag{1}$$

The 10% normative threshold (NT10) methodology, recommended by Wise and Ma (2012), was used to determine which behavior each student displayed according to the response time. After calculating all $SB_{ij,}$ we computed RTE indices as a summary measure of effort for a test (Wise & Kong, 2005). More precisely, the RTE indicates the percentage of items in which an examinee demonstrates solution behavior. As denoted in Equation 2, the overall RTE index for examinee *j* on the test is calculated as follows:

$$RTE_j = \frac{\sum SB_{ij}}{k}, \tag{2}$$

where *k* is equal to the number of items in the test. RTE scores range between 0 and 1, reflecting the proportion of test items for which the examinee demonstrated SB. Consequently, higher RTE values suggest that the examinee likely approached the test items with sufficient effort, while lower RTE values indicate a lack of substantial effort from the examinee (Setzer et al., 2013). For interpreting the behavior type of each grader, the RTE score was divided into three categories: high effort (above .90), medium effort (between .90–.80) and low effort (below .80) in this study (Wise & Kong, 2005). Then, robustly, the grader in the low effort category was acknowledged as having rapid-guessing behavior, and the ones in the high effort category had SB.

2.3. Analysis Procedures

The Statistical Package for the Social Science (SPSS) version 24.0 (IBM Corp., Armonk, NY) program was used for the item-level analysis. First, mean RTs for content and cognitive domains were determined for each item. In the TIMSS 2019 data, there also are many items that were combined, or derived, for scoring purposes, which are called derived items. In our study, the RTs for the derived items were calculated by dividing the number of items contained in these items. Then, the difficulty of each item was calculated. The factorial analysis of variance (ANOVA) was conducted used to determine whether the mean RTs differed according to cognitive domain, content domain, and difficulty level. To determine which categories of these variables differed significantly from each other, the Scheffe test was used, which is one of the post hoc tests. In terms of the second research question, software named the International Association for the Evaluation of Educational Achievement (IEA) International Database Analyzer (IDB) (IEA, 2017) was used to conduct multiple regression analysis, sampling design, sampling weights and plausible values should be considered when analyzing large-scale assessments such as the TIMSS to avoid biased results. With the IDB Analyzer, which made this possible, student total weights (TOTWGT) were used in student level analyses. With the IDB Analyzer, SPSS syntax that considers the sampling weights was generated and multiple regression analysis was performed on SPSS using this syntax. SLM, SCM, HRL and SEC were continuous predictor variables and gender was a dummy-coded predictor variable where girls were the reference group. If the absolute value of the *t*-test is greater than 1.96, the result can be regarded as statistically significant ($p < .05$). Therefore, significance tests are conducted by *t*-value. Partial eta squared (η^2) effect sizes were calculated to determine the proportion of

unique variance of each variable in the analysis. The effect sizes were interpreted using the following benchmarks given by Cohen (1988): small (.01), medium (.06), and large (.14).

3. FINDINGS

3.1. Findings for the First Research Question

First, for item-level analysis, descriptive statistics of mean RT by item difficulty, cognitive domain, and content domain are summarized in Table 2.

Table 2. *Descriptive statistics of mean response time by item difficulty, cognitive domain, and content domain.*

		Item Difficulty							
		Hard		Moderate		Easy		Total	
Country	CD by CnD	M	SD	M	SD	M	SD	M	SD
Singapore	Knowing								
	Numbers	-	-	-	-	37.43	17.21	37.43	17.21
	M & G	-	-	40.64	1.48	38.03	13.37	38.46	12.18
	Data	-	-	-	-	64.99	13.81	64.99	13.81
	Total	-	-	40.64	1.68	41.68	18.34	41.63	17.85
	Applying								
	Numbers	66.14	-	67.98	35.40	52.64	16.99	53.83	17.78
	M & G	-	-	55.45	24.86	65.97	37.25	64.11	34.95
	Data	-	-	112.51	-	74.64	27.30	77.55	28.17
	Total	66.14	-	69.14	31.42	60.15	26.26	61.04	26.42
	Reasoning								
	Numbers	207.94	-	128.77	48.96	64.69	28.72	115.15	57.40
	M & G	102.85	68.26	73.01	20.39	47.01	14.99	64.55	38.03
	Data	182.08	-	138.52	-	79.71	56.80	99.86	61.90
	Total	139.72	70.44	115.64	47.64	61.84	37.48	91.41	55.08
	Total	127.46	69.80	91.64	48.82	52.82	26.75	60.77	37.38
Türkiye	Knowing								
	Numbers	66.47	29.57	50.63	13.77	53.14	23.94	53.70	21.73
	M & G	48.21	2.59	52.57	11.15	53.15	22.27	51.83	14.97
	Data	117.92	-	85.71	5.67	75.57	17.33	84.67	18.54
	Total	63.77	28.51	56.89	17.16	56.13	23.44	57.45	22.08
	Applying								
	Numbers	70.97	30.37	73.28	20.42	68.76	20.45	71.69	21.64
	M & G	72.74	33.70	94.39	34.97	42.79	11.51	81.47	36.50
	Data	114.15	50.04	114.88	46.41	88.04	36.68	106.45	42.60
	Total	82.21	38.11	85.96	33.09	68.71	26.90	80.91	32.93
	Reasoning								
	Numbers	140.83	45.50	106.94	51.31	46.24	-	125.73	51.03
	M & G	94.66	46.19	64.38	26.51	61.57	5.67	74.79	34.93
	Data	148.52	26.50	71.84	6.44	102.14	67.04	106.16	53.77
	Total	127.36	47.38	76.26	34.86	82.57	53.62	100.88	50.04
	Total	98.18	48.25	76.27	32.06	63.52	30.53	76.90	37.89

Note. CD = Cognitive Domain, CnD = Content Domain, M = Mean, SD = Standard Deviation, N = Item Numbers, M & G = Measurement and Geometry.

Table 2 displays the mean and standard deviation of mean RT by item difficulty, cognitive domain, and content domain. Whether the differences observed in Table 2 were statistically significant or not was examined by factorial ANOVA and the main and interaction effects on mean RT are presented in Table 3.

Table 3. *Factorial ANOVA of mean response times by cognitive domain, content domain and item difficulty.*

Country	Source	*df*	MS	F	η^2	Difference
Singapore	CD	2	6860.14	9.72**	.12	K<A<R
	CnD	2	8847.86	12.54**	.15	N<D, M&G<D
	P	2	9168.87	12.99**	.16	E<M<H
	CD x CnD	4	825.23	1.17	-	
	CD x P	3	2762.13	3.91**	.08	
	CnD x P	4	1723.34	2.44*	.07	
	CD x CnD x P	2	128.35	.18	-	
	Error	139	705.66			
	R^2 = .56; adj R^2 = .50					
Türkiye	CD	2	4647.91	5.18**	.07	K<A<R
	CnD	2	9912.91	11.05**	.14	N<D, M&G<D
	P	2	6576.03	7.33**	.10	E<M<H
	CD x CnD	4	550.11	.61	-	
	CD x P	4	3058.55	3.41*	.09	
	CnD x P	4	812.12	.91	-	
	CD x CnD x P	8	938.90	1.05	-	
	Error	132	897.09			
	R^2 = .48; adjR^2 = .38					

Note. CD = Cognitive Domain, CnD = Content Domain, P = Item Difficulty, MS = Mean squares, η^2 = Effect Size, K = Knowing, A = Applying, R = Reasoning, N = Numbers, M & G = Measurement and Geometry, D = Data, H = Hard, M = Moderate, E = Easy, R^2 = .556 and adj R^2 = .495 for Singapore, R^2 = .478 and adj R^2 = .375 for Türkiye, *p < .05. ** p < .01.

As shown in Table 3, all main effects (the cognitive domain ($F_{Singapore}$ (2, 139) = 9.72, p < .01; $F_{Türkiye}$ (2, 132) = 11.05, p < .01), content domain ($F_{Singapore}$ (2, 139) = 12.54, p < .01; $F_{Türkiye}$ (2, 132) = 11.05, p < .01) and the item difficulty ($F_{Singapore}$ (2, 139) = 12.99, p < .01; $F_{Türkiye}$ (2, 132) = 7.33, p < .01), is significantly affected on the mean RT for both samples. According to the Scheffe test for both samples, the source of the differences was the mean RT increased from knowing to reasoning, from easy to hard, and the mean RT of the data content area was higher than that of the other content areas (see Table 2). In terms of two-way interaction, there was only a significant interaction between cognitive domain and item difficulty in the Türkiye sample, and besides a significant interaction between content domain and item difficulty in the Singapore sample, as well. Three-way-interactions did not statistically affect the mean response time. With a large effect size, the highest proportion of the variance of the mean response time in the Singapore sample was attributed to item difficulty, content domain and cognitive domain, respectively, whereas, in the Turkish sample, they were content domain, item difficulty, and cognitive domain, respectively.

3.2. Findings for the Second Research Question

The findings related to the prediction of the mean RT of the items in the mathematics achievement test according to the student-level variables are given in Tables 4 and 5.

Table 4. *Multiple regression results by content domain.*

Variables	Country											
	Singapore						Türkiye					
	Number		M & G		Data		Number		M & G		Data	
	B	β	B	β	B	β	B	β	B	β	B	β
HRL	-1.45	**-.11**	-.52	**-.04**	-.22	-.01	-1.48	**-.13**	-1.04	**-.09**	-1.64	**-.08**
SCM	-.95	**-.09**	.23	.02	.21	.01	-.78	**-.08**	.72	**.07**	.52	.03
SEC	-.75	**-.07**	-.74	**-.06**	-1.20	**-.07**	-.30	-.03	-.36	-.03	.54	.03
SLM	.39	**.04**	.40	**.04**	.48	.03	.95	**.08**	-.07	-.01	.67	.03
Gender[a]	-4.01	**-.10**	-2.67	**-.06**	-5.71	**-.09**	-4.79	**-.11**	-2.13	**-.05**	-2.26	-.03

Note. [a] Girl = 1, Boy = 2. Significant standardized weights ($p < .05$) are bold. HRL: Home Resources for Learning, SCM: Students Confident in Mathematics, SEC: Self-Efficacy for Computer Use, SLM: Students Like Learning Mathematics, M & G = Measurement and Geometry

Table 4 displays the outcomes of the multiple regression analyses conducted on the content domain. The noteworthy negative β weights associated with each predictor variable reveal that students who achieved higher scores in these variables exhibited reduced mean RTs during the TIMSS 2019. Conversely, the significant positive β weights for each predictor variable indicate that students with higher scores in these variables demonstrated increased mean RTs in the TIMSS 2019. Besides, when the results for the gender variable, which is a categorical variable, were negative, it was determined that the RTs of girls were longer than that of boys. But these variables explained 5% of the variance of the mean RT for Singapore (R^2 = .05) and 4% for Türkiye (R^2 = .04). For the number content domain, similar results in both countries for five independent variables were obtained to be significant result (standardized β weight ranges from -.13 to .08), only not for Türkiye for SEC.

For the measurement and geometry content domain, all variables without SCM significantly predicted the mean RT (standardized β weight ranges from -.06 to .04) for the Singapore sample. For Türkiye, only two of the five variables (SEC and SLM) had a non-significant effect on predicting mean RT (standardized β weight ranges from -.09 to .07). But these variables explained only 1% and 2% of the variance of mean RT for Singapore (R^2 = .01) and Türkiye (R^2 = .02), respectively.

For the data content domain, only SEC and gender variables were found significant for Singapore sample (standardized β weight ranges from -.09 to -.07) and only HRL had a significant influence for the Türkiye sample (standardized β weight -.08). But these variables were a part of the variance of the mean RT only with 1% for both samples (R^2 = .01).

Table 5. *Multiple regression results by cognitive domain.*

Variables	Country											
	Singapore						Türkiye					
	Knowing		Applying		Reasoning		Knowing		Applying		Reasoning	
	B	β	B	β	B	β	B	β	B	β	B	β
HRL	-1.24	**-.13**	-1.41	**-.12**	.79	**.03**	-1.35	**-.14**	-2.33	**-.21**	.79	.04
SCM	-.74	**-.10**	-.90	**-.10**	1.15	**.06**	-.40	-.05	-.74	**-.07**	1,76	**.10**
SEC	-.69	**-.09**	-.81	**-.08**	-1.04	**-.05**	-.47	**-.05**	.13	.01	-.57	-.03
SLM	.40	**.05**	.45	**.05**	.44	.02	.40	.04	.56	**.05**	1.14	**.06**
Gender[a]	-2.83	**-.10**	-1.82	**-.05**	-9.26	**-.12**	-1.46	**-.04**	-3.22	**-.08**	-6.53	**-.09**

Note. [a] Girl = 1, Boy = 2. Significant standardized weights ($p < .05$) are bold. HRL: Home Resources for Learning, SCM: Students Confident in Mathematics, SEC: Self-Efficacy for Computer Use, SLM: Students Like Learning Mathematics

Table 5 displays the results of the multiple regression analyses for the cognitive domain. For the knowing domain, all variables were significant predictors (standardized β weight ranges from -.13 to .05) and shared 5% of the variance of the mean RT for Singapore (R^2 = .05). For Türkiye, only two of the five variables (SEC and SLM) had a non-significant effect on predicting mean RT (standardized β weight ranges from -.14 to -.04) and the explained variance was R^2 = .05.

For the applying domain, in Singapore sample, five independent variables were obtained to be significant results (standardized β weight ranges from -.12 to .05). In Türkiye sample, SEC did not have significant standardized β *coefficient (.01; p > .05)*. These variables explained 5% of the variance of the mean RT for Singapore (R^2 = .05) and 4% for Türkiye (R^2 = .04).

For the reasoning domain, all variables without SLM variable significantly predicted the mean RT (standardized β weight ranges from -.12 to .06) for the Singapore sample. For Türkiye, only two of the five variables (SEC and HRL) had a non-significant effect on predicting mean RT (standardized β weight ranges from -.09 to .10). But these variables explained only 2% and 3% of the variance of mean RT for Singapore (R^2 = .02) and Türkiye (R^2 = .03), respectively.

3.3. Findings for the Third Research Question

The findings of the RTE of the students in Singapore and Türkiye 4th grade samples in the mathematics achievement test are presented in Table 6.

Table 6. *Percentage of response time effort categories by content and cognitive domains.*

Domain		Singapore			Türkiye		
		Low	Medium	High	Low	Medium	High
Content	Numbers	.15	.65	99.20	.35	1.02	98.63
	Measurement and Geometry	.15	1.37	98.48	.30	2.04	97.66
	Data	.64	.38	98.96	2.14	.72	97.07
Cognitive	Knowing	.20	1.50	98.40	.30	2.80	97.00
	Applying	.10	2.00	97.90	.50	3.60	96.00
	Reasoning	.30	1.34	98.36	.60	2.48	96.92

Note. High effort (above .90), medium effort (between .90–.80) and low effort (below .80)

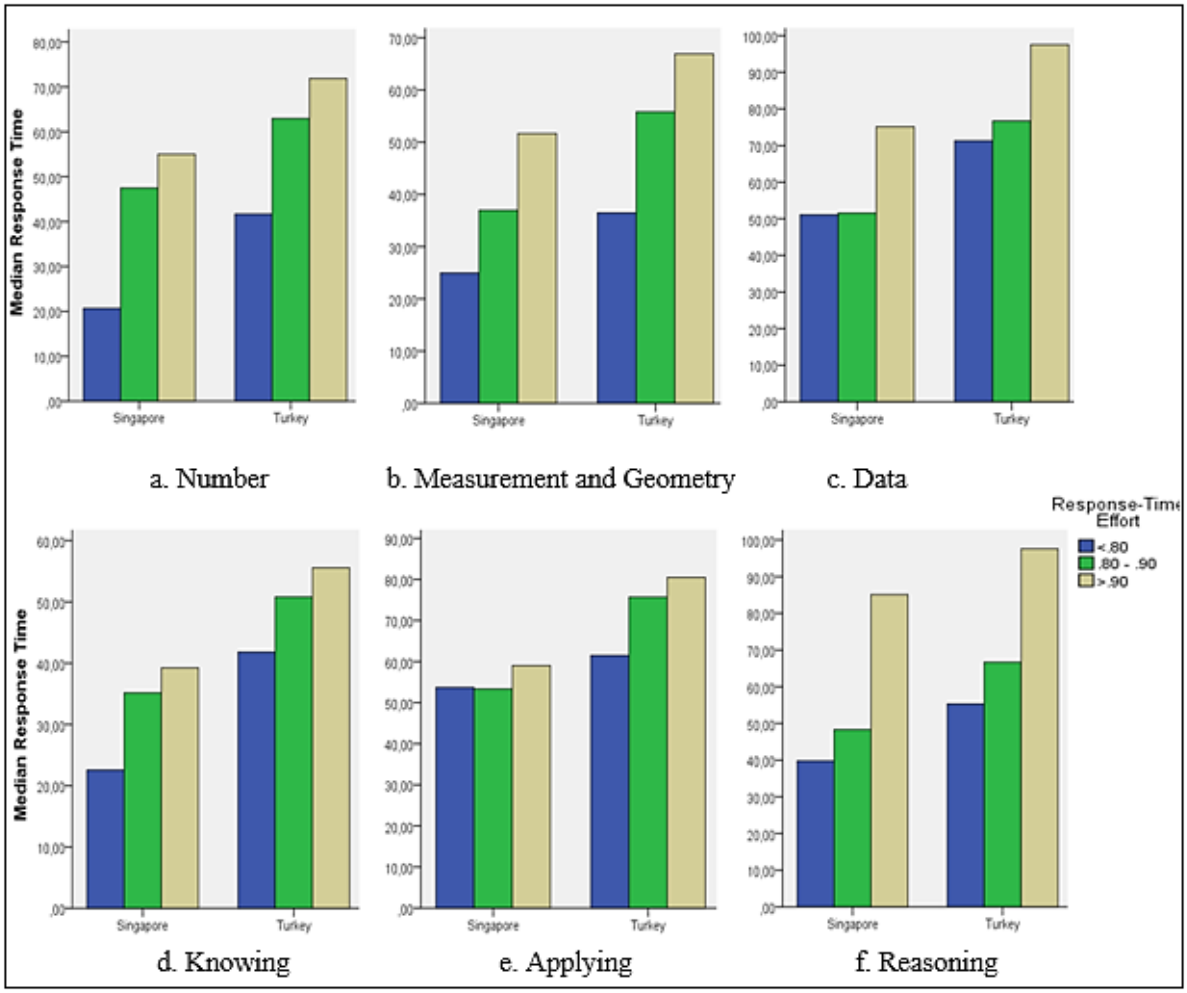
Table 6 presents the percentage of RTE categories by content and cognitive domains. The RTE index for a student was determined by calculating the average of the SB index across the items attempted by the student for each domain. More than 97% of Singaporean students and more than 96% of Turkish students had an RTE value between .9–1.0, which means those students were consistently classified as exhibiting SB across all domains throughout the test. There was a maximum .65% of Singaporean students and .60% of Turkish students (except for the data content domain, it was 2.14%) with RTE smaller than .8, which means those students are more likely to be categorized as displaying rapid guessing behavior.

Only among students with low RTE, when we examined the distribution of categories of HRL, SCM, SLM and SEC, nearly more than half of the students were in the lower category, such as low self-efficacy. The distribution by gender was balanced, that is, the tendency of male and female students to exhibit rapid-guessing behavior was similar.

The mean RT according to RTE categories for each content and cognitive domain is presented in Figure 1. As seen, the graphs for both content and cognitive domains support the assumption that students with a high RTE score consume more energy to get a good score on the test. The reason why we use the median instead of the mean in the graphs is to show the relationship between test-taking performance and RT more clearly. Because the SB index is calculated based

on the threshold, if we used the mean for graphs, it might not show the expected results due to extreme values. For example, if the threshold is 7 seconds, the student who answers an item in 1500 seconds is coded as 1, and another student who answers the same item in 8 seconds is also coded as 1. Here, the mean as the center of gravity is not valid for the distribution of RT and accordingly, the median is appropriate.

Figure 1. *Median response time according to response-time effort by content and cognitive domains.*



4. DISCUSSION and CONCLUSION

4.1. Relationship Between Content Domain, Cognitive Domain, and Item Difficulty with Mean Response Time

The results of this study showed that students' difficulty with the items was positively related to the cognitive level of the item in general. In other words, items from the knowing domain to the reasoning domain became increasingly difficult and the correct answer rate of the students decreased. Certainly, this is a predictable result as the "knowing" aspect encompasses the factual knowledge, concepts, and procedures that students are expected to be acquainted with. However, "reasoning" represents a higher-level cognitive domain that extends beyond solving routine problems, incorporating unfamiliar scenarios, intricate contexts, and multi-step challenges (Mullis & Martin, 2017, p.22). This result supports the postulate of a cumulative hierarchy of the cognitive domains. In this study, this fact was more prominently observed for Turkish students and agrees with the results of many studies on different subject areas (math or science etc.) or data sets (TIMMS or PISA etc.) as well. This result is similar to other studies

showing that the difficulty level of an item will increases according to the cognitive level (Ardıç & Soysal, 2018; İlhan et al., 2020; Koçdar et al., 2016; Nevid & McClelland 2013, Veeravagu et al., 2010). Additionally, Nehm and Schonfeld (2008), Momsen et al. (2013), İlhan et al. (2020), Ardıç and Soysal (2018) stated that item difficulty is not only affected by cognitive level but also by factors such as item type, content, and subject area of the item. In that regard, this study was similarly demonstrated that students, especially Türkiye sample, experienced more difficulties in measurement and geometry, data and numbers in content domains, respectively.

Another finding from this study showed that the content and cognitive domain of the items was positively related to the mean RT. In another word, when the cognitive level of the items increased, both Turkish and Singaporean students spent more time on the solution. Similarly, Yalçın (2022) determined that the cognitive level of the items caused a significant difference on the RTs of the students. Additionally, there were statistically significant differences between the content domains of the item in the mean RTs in both samples. Both Turkish and Singaporean students took longer to respond to data domain items compared to number and measurement and geometry domain items. Lee and Jia (2014) also examined RTs using the 8th grade mathematics items of the National Assessment of Educational Progress (NAEP). Although they found that none of the content areas (algebra, data, geometry, measurement, and number) caused particularly low RTs, the highest median RT was obtained from the number content area. This is different from the relationship between content domain and RT in our study. This difference may be due to the fact that the questions of the exams analysed in the studies were prepared at different cognitive levels. In addition, in the present study, it was determined that as the difficulty of the item increased, the student spent more time on the item. This finding is in parallel with Yang et al. (2002) who found a significant positive relationship between item difficulty and response time. The increase in students' effort while solving difficult items was also observed in the study conducted by Chae et al. (2018).

4.2. The Influence of Home Resources for Learning, Students Confident in Mathematics, Self-Efficacy for Computer Use, Students Like Learning Mathematics and Gender on Mean Response Time

HRL was negatively associated with the mean RT of 4th graders across almost all domains for both countries. This means that when a student has a higher level HRL, he or she will respond in less time to the items, and vice versa. Merely, mean RT in the reasoning domain was positively affected by the HRL score. Reasoning encompasses the application of knowledge and skills to unfamiliar contexts, encompassing the ability to draw logical inferences based on specific assumptions and rules, as well as providing justifications for the obtained results (Mullis et al., 2016, p.24). Therefore, it is an expected finding that students would perform the high-level skills required by items in reasoning in a longer time than items in lower cognitive domains. In this study, the variable with the highest impact on the mean RT was HRL. We think that we contributed to the literature by probably being the first study to examine any relationship between these two variables.

SCM was negatively correlated with the mean RT for items in both the knowing and applying cognitive domains and number content domain but was positively correlated with items in the reasoning cognitive domain and measurement and geometry content domain. This difference may be due to the difference in the difficulty levels of the items according to the content domain. Similarly, Yalçın (2022) found that students who were somewhat confident in mathematics spent less time answering difficult mathematics items than students who were very confident. However, Lasry et al. (2013) stated that students with low self-confidence spent more time to answer the items. In the study by Hoffman and Spatariu (2008), negatively correlation between self-efficacy and RT was found only for easier problems. According to the researchers,

undergraduate students with higher levels of self-efficacy may opt for automatic strategies instead, potentially allocating their time-consuming resources towards problem-solving tasks. Hoffman (2010) also observed similar relationships in his paper with pre-service teachers and ungraduated students. In this study, items in the reasoning domain and items in measurement and geometry domain are also relatively more difficult than ones in other content and cognitive domains. In this context (the two terms are not the same thing, but self-confidence and self-efficacy are so related), our findings are consistent with the paper of Hoffman and Spatariu (2008) and Hoffman (2010).

SLM was positively correlated to mean RT for both samples under most conditions. This variable had no significant effect for items in reasoning from the cognitive domain and in data from the content domain for Singapore sample. But for Türkiye sample, the higher the cognitive domain in which an item was, SLM was significantly more effective on mean RT. This is an unexpected finding because of the positive relationship between self-confidence and like learning mathematics. The students with the confidence, as some researchers reported, will be more motivated and more like learn mathematics (Hannula, 2004; Levine & Donitsa-Schmidt, 1998; Rabbani & Herman, 2017). Additionally, we positively found a correlation with approximately .63 between these variables for both samples. In this study, students who were confident in mathematics and SLM had affected the mean RT in the different way. We think that a self-report bias could affect the emergence of this dilemma. As the American Psychological Association (2022) defines, self-report bias occurs when individuals offer self-assessed measures of some phenomena and individuals may not give answers that are fully correct even if the survey is anonymous. There are many reasons for self-report bias, ranging from a misunderstanding of what a proper measurement is to social-desirability, where the respondent seeks to make a good impression in the survey or not knowing the full answer.

In the Singapore sample, the higher the students had scores on scales of SEC, the sooner they spent time in response. However, this variable did not statistically have any influence on the mean RT in the Türkiye sample. Actually, we found this finding somewhat surprising. Because the ratio of Turkish and Singaporean students who have their own computers is approximately 74% and 95%, respectively, although the scale means of the two countries are quite close. Cultural factors and individual backgrounds could play a role in this phenomenon.

Another finding, there was an influence of gender on the mean RT for both samples. Girls devoted a longer time to response than boys for all domains in the Singapore sample, but for only the number content domain and all cognitive domains in the Türkiye sample. Hunt et al. (2017) analyzed RT data using a 2 (year group, 5 and 6 graders) × 2 (gender) × 2 (problem type, two and three digits) mixed ANOVA. It can be acknowledged that the problems in their study are classified in the number content domain. Unlike our finding, they found that there was no significant main effect of gender and interactions between the independent variables.

4.3. Response Time Effort and Behavior of Students Across the Test

Wise (2017, p.55) interpreted rapid guessing as the following: "Generally, in high-stakes tests, rapid guesses represent strategic attempts to maximize one’s score, whereas in low-stakes tests they represent unmotivated test taking." However, a test taker with SB may have to display rapid guessing if the test with a time limit was about to expire. Irrespective of the reasons and the testing environment, when rapid guessing behavior is observed, it indicates that the test taker is either not engaged or minimally engaged with the test item in terms of effort. In terms of mean RT and RTE index smaller than .8, rapid-guessing behavior and, accordingly, low motivation was observed in below 1% of both samples (except for the data content domain in the Türkiye sample, it was 2.14%). Although it is a low-stakes assessment, almost all the students in the TIMSS 2019 math test showed high SB and, accordingly, high motivation. The variations in students' effort levels on low-stakes tests across different countries could be

attributed to cultural disparities in the significance placed on academic achievement. For instance, Gneezy et al. (2019) found a positive correlation between increased stakes associated with tests and performance in the United States, whereas this correlation was not observed in Shanghai. Borgonovi et al. (2021) highlighted those Asian countries like Singapore place great emphasis on international assessments, considering them as indicators of government policy effectiveness and a source of national pride. This political factor could positively affect the attitudes of Singaporean students toward international tests, and their motivation to do their best. In Türkiye, on the other hand, some studies have been conducted at the provincial and school level to ensure student motivation and to be ready for applications (Ministry of National Education, 2019).

In terms of students classified as low effort in the present study, we observed that Turkish and Singaporean students were likely to have rapid guessing behavior when an item in the reasoning domain became increasingly difficult (probably increasingly complex, also). Similar facts occurred in the data content domain, especially for Turkish graders. Although girls devoted a longer time to the response item than boys, almost no difference was observed in terms of students with low or high RTE index by gender and domains. Only for Singaporean graders, girls had a little higher test-taking effort and test motivation than boys. Unlike this finding, Zhao (2020) reported that girls were less likely to show disengaged behavior than boys in PISA 2012 assessments of computer-based mathematics. Additionally, positive but weak relationships (correlations with maximum .10) were observed between RTE and HRL, SCM, SEC and SLM, which means that graders with higher scores on these scales, he or she would have more items with solution behavior and higher test motivation. Similarly, Zhao (2020) reported a negative correlation between number-disengaged items (refers to rapidly selecting a response to multiple-choice items or omitting items) mathematics interest, and math self-efficiency.

4.4. Limitations and Suggestions for Future Research

In our study, the TIMSS 2019 User Guide (Fishbein et al., 2021) was used for the classification of the cognitive domains in which the items were included. The difficulty of placing the items in a particular cognitive domain precisely can be considered a limitation. For example, some items are likely to belong to more than one cognitive domain, or some experts may disagree on the cognitive categorization of some items. The other limitation of the study, the NT10 methodology was used to display which type of behavior students displayed. But various methods for setting the threshold have been suggested, including the use of mixture modeling (Schnipke & Scrams, 1997), visually inspecting the response time distribution (DeMars, 2007; Wise, 2006), using item characteristics (Wise & Kong, 2005), setting a common threshold across all items (Wise et al., 2004), cumulative proportion method (Guo et al., 2016), mixture log-normal (Rios & Guo, 2020). Therefore, a similar study topic using different threshold methodology can be suggested for future research. Like the study of Walkington et al. (2019), the influence of language features of mathematic problems, such as the number of sentences, pronouns, or problem topics, on student response time could be examined because systematically varying readability may demand affect student performance by different researchers. Since the TIMSS 2019 questions could not be fully accessed, this research was insufficient to examine how the characteristics of the items that need to be examined in person will affect the response time. As another research topic, the effect of students' familiarity and confidence in using a computer or tablet can be examined on test-taking efforts in computer-based test assessment under various conditions, as well.

Declaration of Conflicting Interests and Ethics

The authors declare 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 authors.

Authorship Contribution Statement

Esin Yilmaz Kogar: Problem Statement, Investigation, Methodology, Visualization, Formal Analysis, and Writing-original Draft. **Sumeyra Soysal**: Investigation, Methodology, Visualization, Formal Analysis, and Writing-original Draft.

Orcid

Esin Yimaz Kogar https://orcid.org/0000-0001-6755-9018
Sumeyra Soysal https://orcid.org/0000-0002-7304-1722

REFERENCES

American Psychological Association. (2022). Self-report bias. In APA dictionary of psychology. https://dictionary.apa.org/self-report-bias

Barry, C.L, & Finney, S.J. (2009). *Exploring change in test-taking motivation.* Northeastern Educational Research Association

Barry, C.L., Horst, S.J., Finney, S.J., Brown, A.R., & Kopp, J.P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10,* 342–363. https://doi.org/10.1080/15305058.2010.508569

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: the effects of incentives on motivation and performance. *European Journal of Psychology of Education, 14*, 441–462. http://www.jstor.org/stable/23420343

Bennett, R.E., Brasell, J., Oranje, A., Sandene, B., Kaplan, K., & Yan, F. (2008). Does it matter if I take my mathematics test on a computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6*(9), 1-39. https://files.eric.ed.gov/fulltext/EJ838621.pdf

Bergstrom, B.A., Gershon, R.C., & Lunz, M.E. (1994, April 4-8). *Computer adaptive testing: Exploring examinee response time using hierarchical linear modeling.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA. https://files.eric.ed.gov/fulltext/ED400287.pdf

Borgonovi, F., Ferrara, A., & Piacentini, M. (2021). Performance decline in a low-stakes test at age 15 and educational attainment at age 25: Cross-country longitudinal evidence. *Journal of Adolescence, 92*, 114-125. https://doi.org/10.1016/j.adolescence.2021.08.011

Bridgeman, B., & Cline, F. (2000). *Variations in mean response time for questions on the computer-adaptive GRE General Test: Implications for fair assessment.* GRE Board Professional Report No. 96-20P. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2000.tb01830.x

Chae, Y.M., Park, S.G., & Park, I. (2019). The relationship between classical item characteristics and item response time on computer-based testing. *Korean Journal of Medical Education, 31*(1), 1-9. https://doi.org/10.3946/kjme.2019.113

Chen, G., Cheng, W., Chang, T.W., Zheng, X., & Huang, R. (2014). A comparison of reading comprehension across paper, computer screens, and tablets: Does tablet familiarity matter? *Journal of Computers in Education, 1*(3), 213-225. http://dx.doi.org/10.1007%2Fs40692-014-0012-z

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cole, J.S., Bergin, D.A., & Whittaker, T.A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*(4), 609–624. https://doi.org/10.1016/j.cedpsych.2007.10.002

Cooper, J. (2006). The digital divide: The special case of gender. *Journal of Computer Assisted Learning, 22*, 320–334. https://doi.org/10.1111/j.1365-2729.2006.00185.x

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Wadsworth.

Çokluk, Ö., Gül, E., & Doğan-Gül, C. (2016). Examining differential item functions of different item ordered test forms according to item difficulty levels. *Educational Sciences-Theory & Practice, 16(1)*, 319-330. https://doi.org/10.12738/estp.2016.1.0329

DeMars, C.E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*(1), 23–45. https://doi.org/10.1080/10627190709336946

Eklöf, H. (2007). Test-taking motivation and mathematics performance in the TIMSS 2003. *International Journal of Testing, 7*(3), 311-326. https://doi.org/10.1080/15305050701438074

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education Principles Policy Practice, 17*, 345-356. https://doi.org/10.1080/0969594X.2010.516569

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics, 37*(5), 655-670. http://dx.doi.org/10.3102/1076998611422912

Fishbein, B., Foy, P., & Yin, L. (2021). *TIMSS 2019 user guide for the international database* (2nd ed.). TIMSS & PIRLS International Study Center.

Gneezy, U., List, J.A., Livingston, J.A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: the role of effort on the test itself. *American Economic Review: Insights, 1*(3), 291-308. http://dx.doi.org/10.1257/aeri.20180633

Guo, H., Rios, J.A., Haberman, S., Liu, O.L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*(3), 173-183. https://doi.org/10.1080/08957347.2016.1171766

Hannula. (2004). Development of understanding and self-confidence in mathematics, grades 5-8. *Proceding of the 28th Conference of the International Group for the Psychology of Mathematics Education, 3*, 17-24. http://files.eric.ed.gov/fulltext/ED489565.pdf

Hess, B.J., Johnston, M.M., & Lipner, R.S. (2013). The impact of item format and examinee characteristics on response times. *International Journal of Testing, 13*(4), 295–313. https://doi.org/10.1080/15305058.2012.760098

Hoffman, B. (2010). "I think I can, but I'm afraid to try": The role of self-efficacy beliefs and mathematics anxiety in mathematics problem-solving efficiency. *Learning and Individual Differences, 20*(3), 276-283. https://doi.org/10.1016/j.lindif.2010.02.001

Hoffman, B., & Spatariu, A. (2008). The influence of self-efficacy and metacognitive prompting on math problem-solving efficiency. *Contemporary Educational Psychology, 33*(4), 875-893. https://doi.org/10.1016/j.cedpsych.2007.07.002

İlgün-Dibek, M. (2020). Silent predictors of test disengagement in PIAAC 2012. *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 430-450. https://doi.org/10.21031/epod.796626

İlhan, M., Öztürk, N.B., & Şahin, M.G. (2020). The effect of the item's type and cognitive level on its difficulty index: The sample of the TIMSS 2015. *Participatory Educational Research, 7*(2), 47-59. https://doi.org/10.17275/per.20.19.7.2

Koçdar, S., Karadağ, N., & Şahin, M.D. (2016). Analysis of the difficulty and discrimination indices of multiple-choice questions according to cognitive levels in an open and distance

learning context. *The Turkish Online Journal of Educational Technology, 15*(4), 16–24. https://hdl.handle.net/11421/11442

Lasry, N., Watkins, J., Mazur, E., & Ibrahim, A. (2013). Response times to conceptual questions. *American Journal of Physics, 81*(9), 703-706. https://doi.org/10.1119/1.4812583

Lee, Y.H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*(3), 359–379.

Lee, Y.H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education, 2*(8), 1-24. https://doi.org/10.1186/s40536-014-0008-1

Levine, T., & Donitsa-Schmidt, S. (1998). Computer use, confidence, attitudes, and knowledge: A causal analysis. *Computers in Human Behavior, 14*(1), 125-146. http://dx.doi.org/10.1016/0747-5632(93)90033-O

Lundgren, E., & Eklöf, H. (2020). Within-item response processes as indicators of test-taking effort and motivation. *Educational Research and Evaluation, 26*(5-6), 275-301. https://doi.org/10.1080/13803611.2021.1963940

Martin, M.O., von Davier, M., & Mullis, I.V.S. (Eds.). (2020). *Methods and procedures: The TIMSS 2019 technical report.* The TIMSS & PIRLS International Study Center. https://www.iea.nl/publications/technical-reports/methods-and-procedures-timss-2019-technical-report

Michaelides, M.P., Ivanova, M., & Nicolaou, C. (2020). The relationship between response-time effort and accuracy in PISA science multiple choice items. *International Journal of Testing, 20*(3), 187-205. https://doi.org/10.1080/15305058.2019.1706529

Ministry of National Education (2019, March 19). *Muğla İl Millî Eğitim Müdürlüğü: The TIMSS 2019* [Muğla Provincial Directorate of National Education: TIMMS 2019] https://mugla.meb.gov.tr/www/timss-2019/icerik/2298

Momsen, J., Offerdahl, E., Kryjevskaia, M., Montplaisir, L., Anderson, E., & Grosz, N. (2013). Using assessments to investigate and compare the nature of learning in undergraduate science courses. *CBE-Life Sciences Education, 12*(2), 239-249. https://doi.org/10.1187%2Fcbe.12-08-0130

Mullis, I.V.S., Martin, M.O., Goh, S., & Cotter, K. (Eds.). (2016). *The TIMSS 2015 encyclopedia: Education policy and curriculum in mathematics and science.* The TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/timss2015/encyclopedia/

Mullis, I.V.S., & Martin, M.O. (2017). *The TIMSS 2019 assessment frameworks*. The TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/timss2019/frameworks/

Myers, A.J., & Finney, S.J. (2021). Change in self-reported motivation before to after test completion: Relation with performance. *The Journal of Experimental Education, 89*, 74–94. https://doi.org/10.1080/00220973.2019.1680942

Nehm, R.H., & Schonfeld, M. (2008). Item feature effects in evolution assessment. *Journal of Research in Science Teaching, 48*(3), 237–256. https://doi.org/10.1002/tea.20400

Nevid, J.S., & McClelland, N. (2013). Using action verbs as learning outcomes: Applying Bloom's taxonomy in measuring instructional objectives in introductory psychology. *Journal of Education and Training Studies, 1*(2), 19-24. http://dx.doi.org/10.11114/jets.v1i2.94

Organisation for Economic Co-operation and Development [OECD]. (2015). *Using log-file data to understand what drives performance in PISA (case study), in students, computers and learning: Making the connection*. OECD Publishing. https://doi.org/10.1787/9789264239555-en

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passaged-based tests. *Journal of Technology, Learning, and Assessment, 2*(6), 1–45. https://files.eric.ed.gov/fulltext/EJ905028.pdf

Rabbani, S., & Herman, T. (2017). Increasing Formulate and Test Conjecture Math Competence and Self Confidence in Using the Discovery Learning Teaching Math. *PrimaryEdu: Journal of Primary Education, 1*(1), 119-128. http://dx.doi.org/10.22460/pej.v1i1.488

Rios, J.A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of diferential nonefortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education, 33*(4), 263-279. http://dx.doi.org/10.1080/08957347.2020.1789141

Schnipke, D.L., & Scrams, D.J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*, 213–232. https://psycnet.apa.org/doi/10.1111/j.1745-3984.1997.tb00516.x

Setzer, J.C., Wise, S.L., van de Heuvel, J.R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education, 26*(1), 34–49. https://doi.org/10.1080/08957347.2013.739453

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review, 31*, 100335. https://doi.org/10.1016/j.edurev.2020.100335

Sundre, D.L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and nonconsequential test performance?. *Contemporary Educational Psychology, 29*(1), 6-26. https://psycnet.apa.org/doi/10.1016/S0361-476X(02)00063-2

Swerdzewski, P.J., Harmes, J.C., & Finney, S.J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education, 24*(2), 162–188. http://dx.doi.org/10.1080/08957347.2011.555217

Veeravagu, J., Muthusamy, C., Marimuthu, R., & Subrayan, A. (2010). Using Bloom's taxonomy to gauge students' reading comprehension performance. *Canadian Social Science, 6*(3), 205–212. https://doi.org/10.3968/J.CSS.1923669720100603.023

Walkington, C., Clinton, V., & Sparks, A. (2019). The effect of language modification of mathematics story problems on problem-solving in online homework. *Instructional Science, 47*(5), 499-529. https://link.springer.com/article/10.1007/s11251-019-09481-6

Wang, M. (2017). *Characteristics of item response time for standardized achievement assessments* [Doctoral dissertation]. University of Iowa.

Wang, T., & Hanson, B.A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29*(5), 323-339. https://doi.org/10.1177/0146621605275984

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement, 41*(2), 115-129. https://doi.org/10.1177%2F0146621616676791

Wise, S.L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education, 19*(2), 95-114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S.L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52-61. https://doi.org/10.1111/emip.12165

Wise, S.L., & DeMars, C.E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educational Assessment, 10*(1), 1-17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S.L., Kingsbury, G.G., Thomason, J., & Kong, X. (2004, April 13-15). *An investigation of motivation filtering in a statewide achievement testing program.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Wise, S.L. & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S.L., & Ma, L. (2012, April 13-17). *Setting response time thresholds for a CAT item pool: The normative threshold method*. In annual meeting of the National Council on Measurement in Education, Vancouver, Canada (pp. 163-183). https://www.nwea.org/resources/setting-response-time-thresholds-cat-item-pool-normative-threshold-method/

Wolgast, A., Schmidt, N., & Ranger, J. (2020). Test-taking motivation in education students: Task battery order affected within-test-taker effort and importance. *Frontiers in Psychology, 11*, 1–16. https://doi.org/10.3389/fpsyg.2020.559683

Yalçın, S. (2022). Examining students' item response times in eTIMSS according to their proficiency levels, selfconfidence, and item characteristics. *Journal of Measurement and Evaluation in Education and Psychology, 13*(1), 23-39. https://doi.org/10.21031/epod.999545

Yang, C.L., O Neill, T.R., & Kramer, G.A. (2002). Examining item difficulty and response time on perceptual ability test items. *Journal of Applied Measurement, 3*(3), 282-299.

Zenisky, A.L., & Baldwin, P. (2006). *Using item response time data in test development and validation: Research with beginning computer users*. Center for educational assessment report No, 593. Amherst, MA: University of Massachusetts, School of Education.

Zhao, W. (2020). *Identification and validation of disengagement measures based on response time: An application to PISA 2012 digital math items* [Master's thesis]. University of Oslo.

Zhang, T., Xie, Q., Park, B.J., Kim, Y.Y., Broer, M., & Bohrnstedt, G. (2016). *Computer familiarity and its relationship to performance in three NAEP digital-based assessments* (AIR-NAEP Working Paper No. 01-2016). American Institutes for Research.

International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 194–207
https://doi.org/10.21449/ijate.1407315
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Collaborative problem-solving design in large-scale assessments: Shedding lights in sequential conversation-based measurement

Qiwei He [1*]

[1]Georgetown University, Data Science and Analytics Program, Washington, DC, 20057 USA

ARTICLE HISTORY

Received: Dec. 20, 2023
Accepted: Dec. 24, 2023

Keywords:
Collaborative-problem solving,
Conservation path,
Sequential measurement
PISA,
Item design.

Abstract: Collaborative problem solving (CPS) is inherently an interactive, conjoint, dual-strand process that considers how a student reasons about a problem as well as how s/he interacts with others to regulate social processes and exchange information (OECD, 2013). Measuring CPS skills presents a challenge for obtaining consistent, accurate, and reliable scale across individuals and user populations. The Programme for International Student Assessment (PISA)'s 2015 cycle first introduced an assessment of CPS in international large-scale assessments in which computer-based conversational agents were adapted to represent team members with a range of skills and abilities. This study draws on measures of the CPS domain in PISA 2015 to address the challenges and solutions related to CPS item design and shed lights on sequential conversation-based measurement. Specifically, we present the process of CPS item design, the development of scoring rules through CPS conversation paths, and discuss the possible approaches to better estimate CPS beyond item response models.

1. LANGUAGE MODELS IN AUTOMATED ESSAY SCORING

Researchers consider the importance of collaborative problem solving as an educational outcome and a skill for life and work as having increased since the turn of the 21st century (National Center for Educational Statistics, 2015; National Academies, 2012; Wildman et al., 2012; Casner-Lotto & Barrington, 2006). Noncognitive skills that intersect with cognitive ones now involve mastering new challenges and require cooperative efforts among a group of individuals (Griffin et al., 2012; Greiff et al., 2014). Collaborative problem solving (CPS) is inherently an interactive, conjoint, dual-strand process that considers how the student reasons about a problem as well as how he or she interacts with others to regulate social processes and exchange information (Organisation for Economic Co-operation and Development [OECD], 2013). While measuring CPS skills presents a challenge for obtaining consistent, accurate, and reliable measurement across individuals and across user populations, it is an also opportunity to gain more information about cognitive processes in interactions with peers. (He et al., 2017).

As Stecher and Hamilton (2004) observed, CPS skills are difficult to measure. Challenges persist from two major aspects: first, developing items with complex constrains, and second, producing a reliable scale to measure the CPS skills in an accurate way. Given concerns about

*CONTACT: Qiwei He qiwei.he@georgetown.edu Georgetown University, Faculty of Education, Department of Educational Sciences, Washington, DC, 20057 USA

e-ISSN: 2148-7456 /© **IJATE** 2023

language factors and fairness across different countries and cultures, even more difficulties have to be confronted when measuring CPS skills in large-scale assessments such as the Programme for International Student Assessment (PISA). Traditional methods that have been generally used for item response modeling may not be appropriate for measuring collaborative interactions because of the dependence within elements of complex tasks and between interacting participants (Cooke et al., 2012; Quellmalz et al., 2009). Therefore, new assessment designs and statistical methods that capture the dynamic of knowledge sharing in collaborative contexts are required (Dede, 2012). How to model such knowledge and skills in a way that meets the technical standards of traditional assessments is an issue that urgently needs to be solved.

A bold move was made in PISA 2015 to introduce CPS to the assessment program (OECD, 2013). This objective was accomplished through the successful implementation of conversational agents incorporated in computer-based testing. Such innovation introduced a new perspective to understanding students' performance that goes beyond the borders of domain-specific competencies and mere cognitive ability constructs such as reasoning and working memory (Greiff et al., 2014). Skillfully dealing with new problems in diverse settings and contexts, as part of a team instead of individually, is at the core of the concept of CPS. CPS reflects a set of skills that combines cognitive and social aspects that are relevant for successful problem solving across domains regardless of the specific contextual setting (He et al., 2017).

The triennial PISA study aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-olds. In 2015 over a half million students, representing 28 million across 72 countries and economies, took PISA in three core cognitive domains—science, reading, and math, as well as CPS and financial literacy. PISA has a history of measuring problem-solving skills, specifically individual problem solving in PISA 2003 (paper and pencil based) and 2012 (computer based), acknowledging these skills' increasing relevance.

This study draws on measures of the CPS domain in PISA 2015 to address the challenges and solutions related to CPS item design and shed lights on sequential conversation-based measurement. Specifically, we present the process of CPS item design, the development of scoring rules through CPS conversation paths, and discuss the possible approaches to better estimate CPS beyond item response models.

In the following section, we introduce the process of CPS item design for PISA 2015 and examine factors that potentially make impact on CPS item difficulty. The CPS scoring rules are illustrated through conversation paths in Section 3. We finalized this paper with a general conclusion on reliability of CPS scale in PISA 2015 and some discussions on future research directions for CPS assessments.

2. DEVELOPING CPS ITEMS FOR PISA 2015

2.1. CPS Item Design

For PISA 2015, CPS is defined as follows: "CPS competency is the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution" (OECD, 2013). As such, this competence integrates two essential concepts: problem solving and collaboration, which were categorized into a set of 12 CPS skills. As shown in Table 1, a matrix of CPS skills was created that included three major CPS competencies crossed with four problem-solving processes.

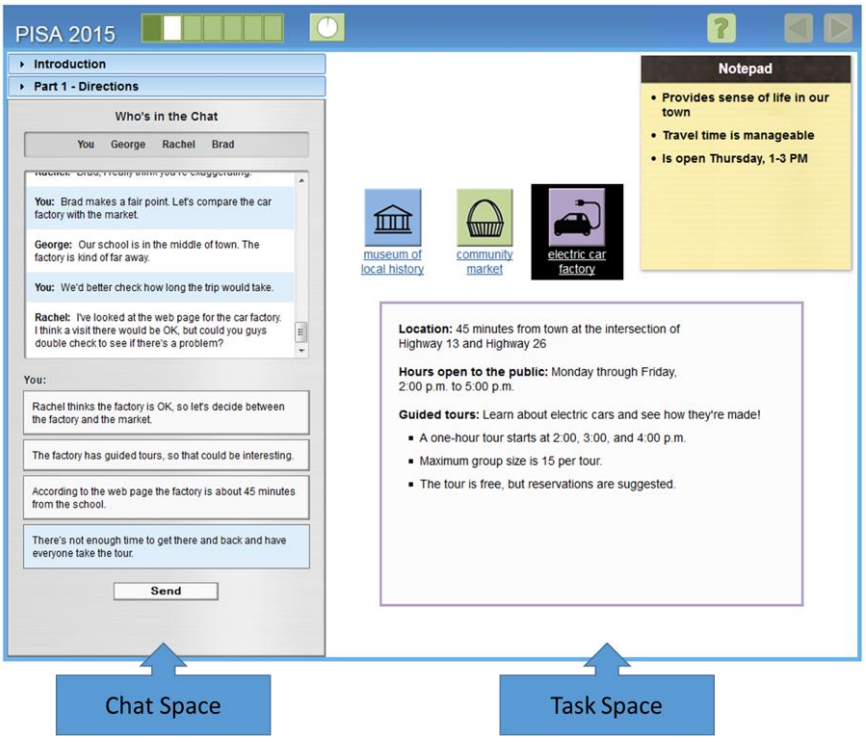
The computer-based CPS tasks (see Figure 1) that were developed to measure these skills were situated in a chat environment ("chat space") where students interacted with one or more simulated agents, identified as teammates, to solve a presented problem. The student was provided with a set of four chat options, and agent responses were based on the option selected.

Each task also included a problem space ("task space") where the student and/or agents could take actions as they worked toward a problem solution. Examples of these actions included selecting information to complete a form or scheduling tasks on a calendar presented in the problem space.

Table 1. *Matrix of Collaborative Problem Solving Skills for PISA 2015 (Organisation for Economic Co-operation and Development, 2013).*

CPS Competency / Problem Solving	(1) Establishing and maintaining shared understanding	(2) Taking appropriate action to solve the problem	(3) Establishing and maintaining team organisation
(A) Exploring and Understanding	(A1) Discovering perspectives and abilities of team members	(A2) Discovering the type of collaborative interaction to solve the problem, along with goals	(A3) Understanding roles to solve problem
(B) Representing and Formulating	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)	(B2) Identifying and describing tasks to be completed	(B3) Describe roles and team organisation (communication protocol and rules of engagement)
(C) Planning and Executing	(C1) Communicating with team members about the actions to be/ being performed	(C2) Enacting plans	(C3) Following rules of engagement, (e.g., prompting other team members to perform their tasks.)
(D) Monitoring and Reflecting	(D1) Monitoring and repairing the shared understanding	(D2) Monitoring results of actions and evaluating success in solving the problem	(D3) Monitoring, providing feedback and adapting the team organisation and roles

Figure 1. *A sample screen of chat and task spaces in a released CPS item (The Visit) in PISA (Organisation for Economic Co-operation and Development, 2015a).*



2.2. Mapping Items onto CPS Skills

As part of the item development process, each item was classified into one of the 12 CPS categories, reflecting the 12 intersecting skills being assessed. Data from the PISA CPS assessment was analyzed to estimate a set of item characteristics for the 117 items included in the main survey.[†] Following data analysis, the items were associated with their difficulty estimates and framework classifications to create an item map. As shown in Table 2, the item map includes information on a certain item along with a brief qualitative description for a subset of CPS items by rows. Table 2 presents two selected items from a released PISA CPS unit (*Xandar*) to illustrate the mapping process, in which the more difficult item is listed first.

Table 2. *A Map for Selected Collaborative Problem-Solving Items in the Released Unit (Xandar).*

Item (Unit and Item ID)	Item Difficulty on CPS Scale	Task Requirements	Establishing & Maintaining Shared Understanding	Taking Appropriate Action to Solve the Problem	Establishing & Maintaining Team Organisation	Exploring & Understanding	Representing & Formulating	Planning and Executing	Monitoring & Reflecting
Xandar CC100203	537	TAKE INITIATIVE by identifying one remaining task needed to solve the problem. Recognize time limits presented in the scenario and assume responsibility for completing the task without further discussion.			○		○		
Xandar CC100301	357	ACT based on agreed-upon role to complete simple assigned task in an uncomplicated problem space.			○			○	

2.3. Examining Factors That Impact CPS Difficulty

The analysis performed to create an item map made it possible to look for factors associated with item difficulty. This could be done by examining the ways in which CPS skills are associated with items located at different points, ranging from the bottom to the top of the scale. When developing a CPS unit, complex constrains may set on items' difficulty level, in order to make a proper mix for items with different difficulty levels. We listed a set of major attributes as below.

2.3.1. *Features of problem complexity*

Features of problem complexity take a high priority in developing a CPS item in accordance with proper difficulty level. There are three major features to help define problem complexity: the nature of the presented problem, the progression of the problem solution, and characteristics of the task space where the problem is worked on.

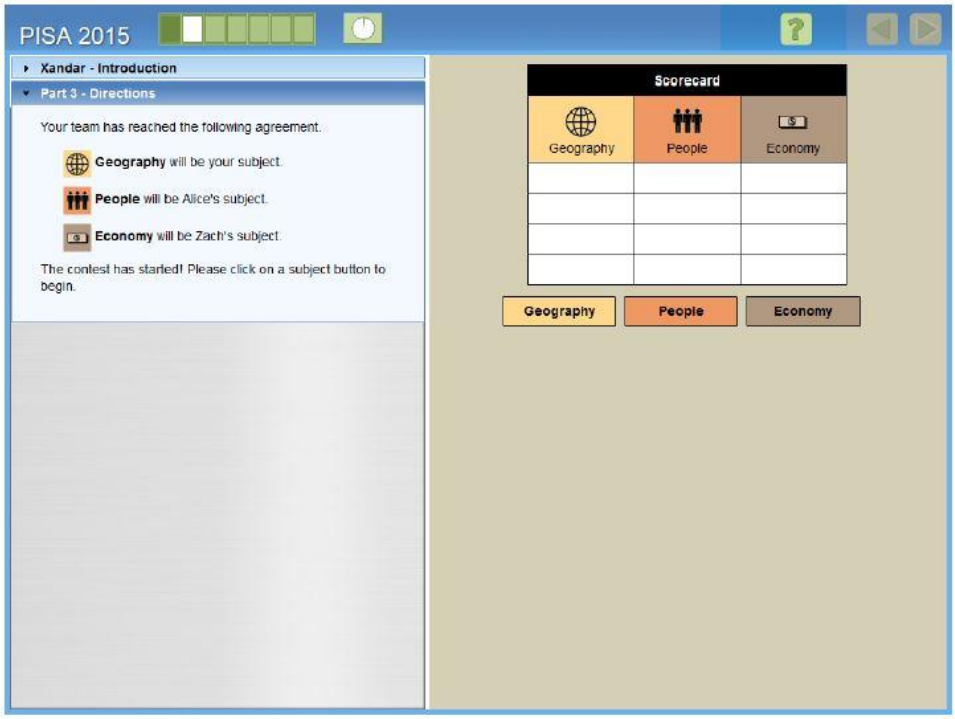
The nature of the presented problem is the first essential element to influence CPS problem complexity. At the lower levels of the scale, problems are well defined with clear goals. Students may be asked to execute a simple and agreed-upon solution, while at higher levels,

[†] This is the number of independently scored items in the final CPS database. Four items included in the main survey were dropped during data analysis. Additionally, a number of items in each unit were combined, based on the main survey analysis and/or to reflect the branching logic within units. As a result of the branching, based on the path students took, students might not see all items in a unit and, therefore, items needed to be clustered in order to function psychometrically.

problems are more complex, requiring students to satisfy multiple constraints, hold more information in working memory, or deal with an impasse or unexpected action. Figure 2 exhibits the screenshot of the lower difficulty item (CC100301) in Table 2. To solve this CPS task, students needed to simply act based on the agreed-upon role, respond to the directions on the screen, and click the correct button. Conversely, the higher difficulty item (CC100203) listed in Table 2, as shown in Figure 3, displayed complexity in the item layout, with an interactive map as well as two tables with dynamic results through the CPS process providing supplementary information. This item required that students respond to a question from one team member and also provided information about how the team is progressing. The additional requirement to identify gaps that had not yet been filled in provides further evidence of its high difficulty level. Students had to use the information displayed in the task space, along with an understanding of how the game worked, to respond correctly.

Figure 2. *A sample CPS item with lower difficulty in a released CPS unit (Xandar) in PISA 2015 (OECD, 2015a).*

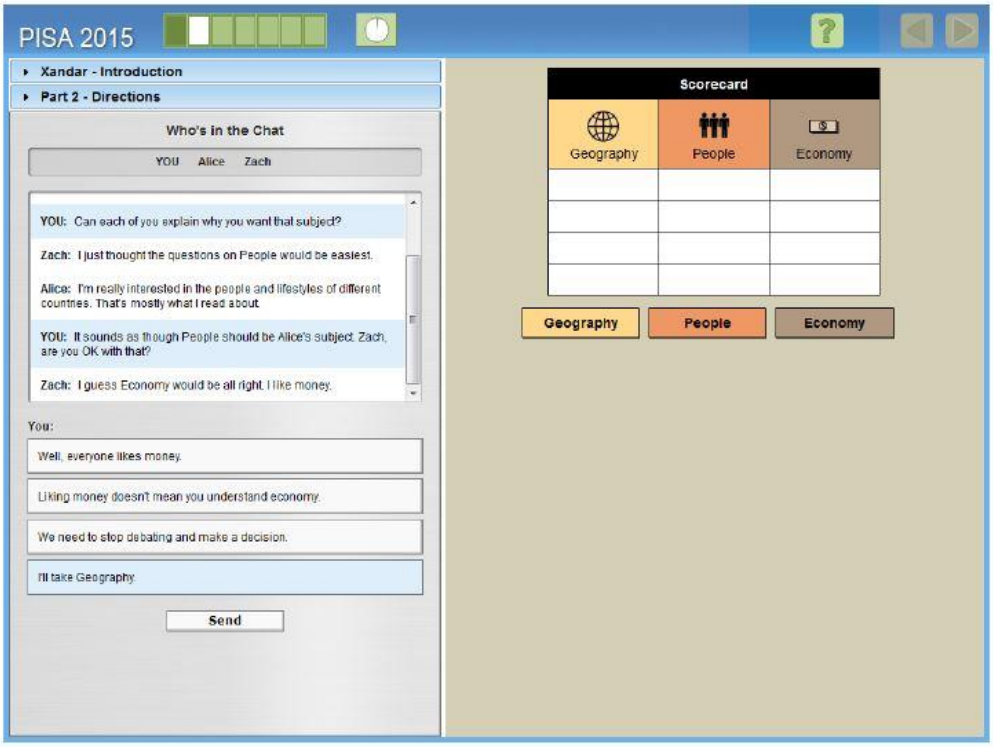
Item	CC100301
Collaborative competency	Establishing and maintaining team organisation
Problem-solving process	Planning and executing
Collaborative problem-solving skill	Following rules of engagement (e.g., prompting other team members to perform their tasks)
Difficulty	357 (Level 1)
Credited action	Student clicks on the "Geography" button.



The second feature of problem complexity relates to the progression of the problem solution. The CPS tasks were chat-based scenarios where information unfolded throughout the course of the task. Item difficulty could therefore be impacted by how recently required information was presented. Having to recall or go back and review information presented earlier in the task makes an item harder to answer. A sequence effect also impacts difficulty in these tasks. Items tend to be easier if they are part of a series of items focusing on a single aspect of the problem and requiring similar student responses.

Figure 3. *A sample CPS item with higher difficulty in a released CPS unit (Xandar) in PISA 2015 (OECD, 2015a).*

Item	CC100203
Collaborative competency	Establishing and maintaining team organisation
Problem-solving process	Representing and formulating
Collaborative problem-solving skill	Describe roles and team organisation (communication protocol/rules of engagement)
Difficulty	537 (Level 2)
Credited response	"I'll take Geography."



Characteristics of the task space are considered the third major feature of problem complexity that influence CPS item difficulty. At lower levels of the scale, changes in the problem space are controlled by the student. Problems may require information to be reordered or new information to be added, but those actions are performed by the respondent. Where information in the problem space changes as a result of agent actions, items tend to be more difficult, particularly in cases where those actions are not explicitly signaled in the chat. In these cases, the student must both notice the changes and infer which of the agents took the action.

Additional aspects of the problem space may affect how difficult it is to solve the presented problem. These include but are not limited to reading load, multiple channels of information-including tables, figures, and diagrams—and the need to use spatial or temporal skills.

2.3.2. *Features of collaboration complexity*

Features of collaboration complexity are often presented by the number of collaborators that are required to be involved in the task and the roles they need to play. In each CPS unit, the student worked with one or more group members to solve a problem, with the group members/computer agents providing input much as fellow students would do. The conversational agents responded to students' textual inputs and actions when the student moved through different stages of the problem. In each stage, communications or actions that could be performed by either the agent or the student was predefined, which resulted in the ability to objectively score all responses.

The computer dynamically monitored the state of the problem through the task completion process. Characteristics of the agents, or virtual team members, with whom the student had to interact also impacted item difficulty. Where agents were collaborative and capable and take an active role in solving the problem, items tended to be easier. In such cases, the student could simply be called upon to provide requested information or agree to the direction being suggested. When agents were focusing on their own goals rather than those of the team, it was more challenging to establish team organization and develop a shared understanding of the problem. The need to collaborate with agents who make errors that need to be noted and remedied can make items more difficult.

The roles of students involved in the collaborative task are also critical to the problem complexity. At the lowest levels of the scale, tasks required only that students respond to agent requests for information or suggestions for actions. More difficult tasks required that students take initiative. That initiative might take several forms including: requesting needed information, suggesting actions for team members to take, and monitoring agent's actions or statements to be sure they are correct and aligned with the agent's agreed-upon role on the team. At higher levels of the scale, tasks required students to resolve a conflict between agents, propose that the team pursue a new approach, or help balance a desire for consensus against the efficiency of the problem-solving process.

2.3.3. *Integration of problem solving and collaboration demands*

Last but not least, the problem solving and collaboration features of each CPS task do not operate in isolation. The difficulty of any given CPS task depends on the interaction between the problem-solving demands and the nature of the collaboration that is required. At the lowest level, items often required either simple collaboration efforts or simple problem-solving tasks. At the highest levels on the scale, complex problem-solving demands were complicated by challenging social interactions, and students had to balance both in order to successfully complete a task.

3. DEFINE CPS SCORING RULE WITH CONVERSATION PATH

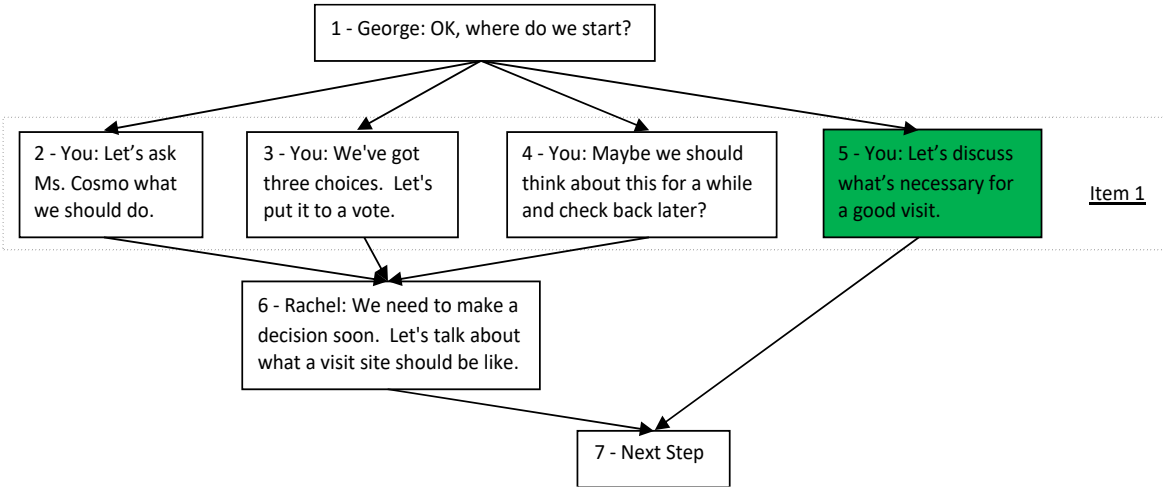
After gaining insight on investigating factors potentially influencing CPS difficulty, we proposed scoring rules for different item types that are specified for the CPS items via path analysis. The construction of different item types was often associated with requirements of different item difficulty levels. To satisfy the conditions of item difficulty level in a specific unit, special item type such as "rescue" and "bonus" items were proposed particularly for the CPS domain. We used some example items here to illustrate how the conversation path analysis worked and how we had to combine items in certain types to keep the CPS scale reliable.

3.1. Conversation Path and Convergence Structure

The major difference between CPS and regular problem solving relates to the perspective of collaboration. In the PISA CPS domain, respondents were required to solve the problem through a collaborative effort, that is, completion of a task with at least two students together rather than an individual alone. As introduced earlier, in one CPS unit, one or many conversational agents worked together with the respondent to go through the dialogues and make "joint" efforts to solve a task. Similar to a computer game, a CPS unit required the respondent to choose an optimal sentence from a set of multiple choices to go through the conversation with agents or choose one or more actions to pass.

Convergence was generally used to guarantee that different paths arrived at an identical point. That is, regardless of what choices the student made, the path led to the same convergence point. Each path to the convergence point had to provide the student with the same information and bring him or her to the same stage of the problem.

Figure 4. *Conversation path of a typical example of CPS item with a simple segment in a released CPS unit (The Visit) in PISA 2015.*



Note. The node highlighted in green is the correct response to this item.

Figure 4 shows an example item with a simple convergence structure in a released CPS unit (The Visit). The collaboration task in this unit was to jointly create a welcoming activity for students coming from abroad. "You" in the script represented the respondent who was required to work with three fellow students (agents)—George, Rachel, and Brad (who shows up later)—to decide what to do to welcome the foreign student. After seeing the input from George (Node 1), the test taker could choose one answer among Node 2 to Node 5 (i.e., Item 1). The respondent got full credit when Node 5 was selected (green). The path then continued to Node 7: the final convergence point. Otherwise, Rachel's response (Node 6) would appear as an intermediate point, and then the path would move on to Node 7 the final convergence point. We defined the phase between two convergence points as one segment, meaning only two convergence points could be found within one segment, the starting convergence point and ending convergence point. A simple segment could have only one scoring item, while a complex segment could have more than one scoring items.

3.2. "Rescue" Items

"Rescue" items were typically developed in a complex segment, where respondents might have the possibility of going through two or three choice points before getting to the convergence point. For example, in The Visit unit, the student and the agents needed to help one of the foreign students get to the airport (see Figure 5 for the item screen and Figure 6 for the conversation path map). The full credited response was the third choice ("I'm at school, where are you guys?"), that is, Node 80 in Figure 6, which told the team his or her location and led directly to the convergence point (Node 85). But students who chose the other paths still arrived at the convergence point, although it took longer. For instance, if the student selected the first option ("What happened to his host family?"), that is, Node 81, Rachel rescued by saying she didn't know what happened to his host family and asking the student if he or she were at school; this gave the student a second chance to choose the response providing his or her location (in Node 83, Node 87, Node 88, and Node 89). If the student selected the second option ("You're good at arranging things, Rachel, can you take care of Zheng?"), that is, Node 78, or the fourth option ("I'm not sure I'm the best person to decide. Rachel, can you help Zheng?"), that is, Node 79, the conversation path worked its way to the final convergence point, meaning students choosing the second and fourth options would not have a chance to answer Item 3. It was noted that the process data in the log file indicated students were unlikely to notice these convergence

and rescue structures. The structure design apparently had little impact on students' test-taking behavior as they progressed through the scenario.

Figure 5. *A sample screen of rescue designs in a released CPS item (The Visit) in PISA (OECD, 2015a).*

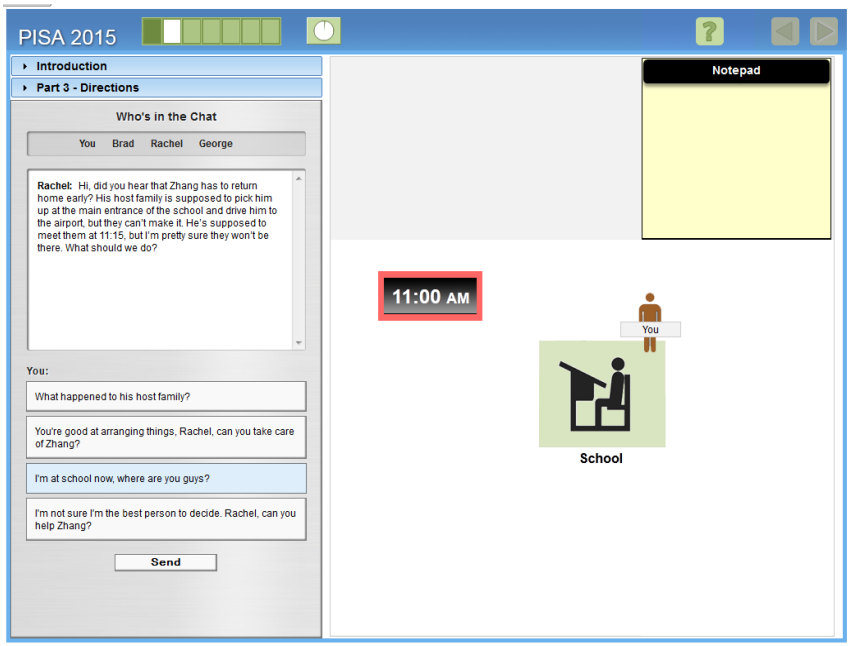
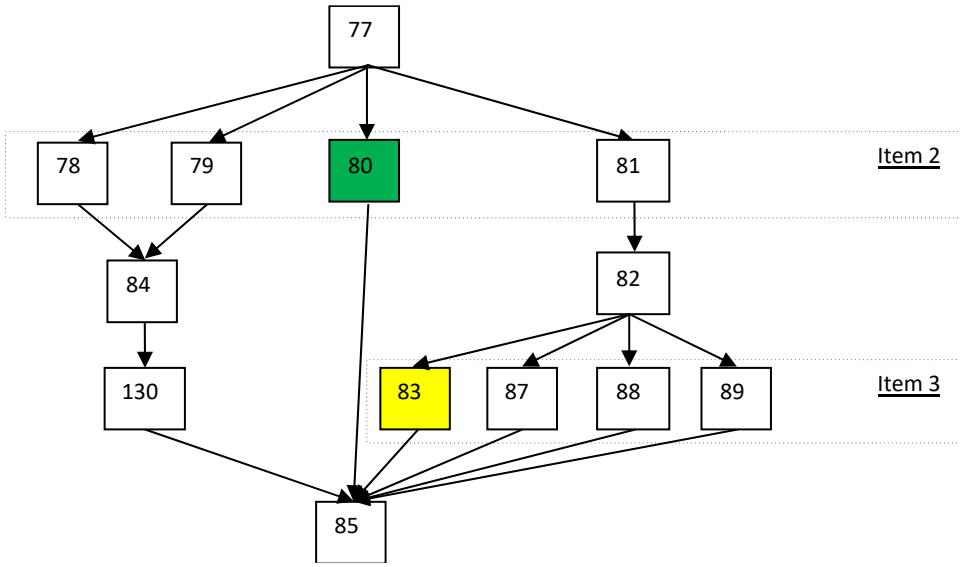


Figure 6. *Conversation path of a "rescue item type" example (see item screenshot in* Figure 5*) of CPS item with a complex segment in a released CPS unit (The Visit) in PISA.*



Note. The node highlighted in green is the correct response to Item 2; the node highlighted in yellow is the correct response to Item 3 on the "rescue" path.

However, the "rescue" item type brought an issue in scoring. Students who got a full credit of 2 points in Item 2 lost the chance to see Item 3, so their Item 3 score was 0; students who got 1 point in Item 3 had already failed in Item 2, recorded as 0 points in Item 2. Therefore, the score correlation between these two items could be substantially negative. One possibility would have been for students who did not have a chance to see Item 3 to receive a score of "not applicable," but such a solution ran counter to the design purpose to assess students' CPS skills based on the selection to the prompt in Node 77.

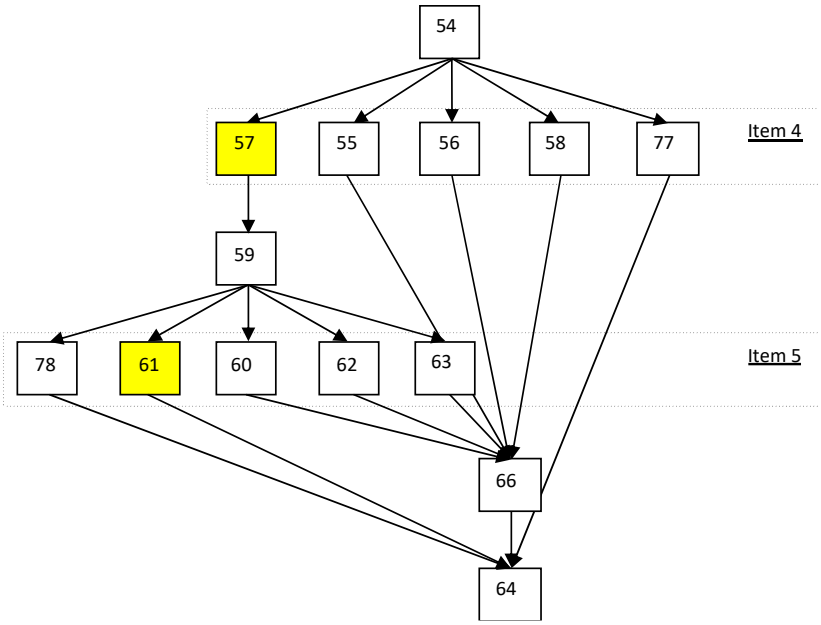
A better solution for such a scoring issue was adopted: to treat the whole complex segment as one polytomous item. Basically, we assigned all item credits within the segment with proper weights. Instead of looking at the individual items, we gave the credit scoring rule in the larger segment, namely, in the combined polytomous item (Item 2 + Item 3): When the test taker's path went through Node 80, the score was 2; when the path went through Node 83, the score was 1; otherwise, the score was 0.

Moreover, it was noticed that even though a single item in a complex segment had already been designed as a polytomous item, we still could transform the segment into a bigger polytomous item by adding up all scores across items and setting full credit, partial credit, and no credit according to the paths.

3.3. "Bonus" Items

Alternatively, a "bonus" item type could also be present in a complex segment. As the path map shows in Figure 7, students who got a full credit of 1 point in Item 4 (Node 57) had an additional chance to score another point in the subsequent Item 5 (Node 61), while the students who answered incorrectly in Item 4 lost the chance of getting a point in Item 5. The point in Item 5 was a "bonus" for students who gave a correct response in Item 4. Considering that the correlation between Item 4 and Item 5 had a very small chance to be negative, we did not put such "bonus" segments into a polytomous item combination.

Figure 7. *Conversation path of a "bonus item type" example of CPS item with a complex segment in a released CPS unit (The Visit) in PISA 2015.*



Note. The nodes highlighted in yellow are the correct responses to Item 4 and Item 5 respectively.

4. DISCUSSION

Collaborative problem solving is a critical competency in a variety of contexts, including the workplace, school, and home. With the increasing growth of digital tasks, collaborations are not only conducted in real practice but also in the virtual environment. As Dede (2009) has observed, "The nature of collaboration is shifting to a more sophisticated skillset. In addition to collaborating face-to-face with colleagues across a conference table, 21st century workers increasingly accomplish tasks through mediated interactions with peers halfway across the world whom they may never meet face-to-face. … Collaboration is worthy of inclusion as a

21st century skill because the importance of cooperative interpersonal capabilities is higher and the skills involved are more sophisticated than in the prior industrial era."

With the debut of CPS assessment in PISA, it is important to prepare a proper measure to keep the CPS scale reliable and valid. the PISA 2015 CPS units were based on simulated conversations with one or more computer-based agents that were designed to provide a virtual collaborative problem-solving situation. Test takers had to choose an optimal sentence from a multiple-choice list to go through the conversation with agents, or choose one or more actions programmed in the unit. Because of the similar item structures in other domains in PISA 2015, the data collected in the CPS units were evaluated by IRT models (Lord, 1980; Rasch, 1960)—specifically, the two-parameter-logistic model and the generalized partial credit model—to establish reliable, valid, and comparable scales. Readers can refer to the PISA 2015 technical report for the details about scaling and analytic procedures (OECD, 2017). The CPS scale in the main survey consisted of six units, which in turn comprised multiple items within each unit that can be used for the IRT scaling. It was found that data from two units had dependencies in the responses due to different paths that could be taken by students through the simulated chat as a result of the "rescue" item type. Therefore, the CPS chat items that showed this kind of dependency were combined into "composite items" by summing the responses for the different paths that respondents could take. With this approach, it was determined that each path-based response string could be scored to provide valid data and introduced into the IRT analysis. The composite items were used to generate polytomous items for the purpose of reducing issues with local dependencies.

To ensure the computational models were used in an appropriate way, we combined items with high correlations by two steps: first, based on the conversation path analysis, each segment with the "rescue" item type was combined into a polytomous item; and second, the remaining items that still had high correlations in the residual analysis were further combined into a "super" item in the latent trait estimation. This approach is superior in standard large-scale assessments to keep consistent with the whole measurement framework across countries. According to the PISA 2015 tech report, the residuals in CPS domain were under a good control and the local dependency of combined super items were well adjusted.

However, the CPS item design proposed a new challenge in sequential conversation-based measurement. Because of the inherent relationship in conversations, the local independency may not adapt to the assumptions of item response models. Given concerns on the dynamic dependency of at least one previous conservation (or even more), the sequence of the conversation path through the whole unit, i.e., vertical measurement path may be given different difficulty parameters rather than each checkpoint on the conversation line, i.e., horizontal measurement by each item, which the local independency has to be assumed but might not be completely correct.

In addition, the CPS framework with computer agents was compatible with the capabilities of the PISA 2015 computer platform. The student could interact with the agents via a chat window, allowing the student to respond through communication menus. With respect to the student inputs, there were conventional interface components, such as mouse clicks, sliders for manipulating quantitative scales, drag and drop, cut and paste, and typed text input. Aside from communicating messages, the student could also perform actions on other interface components. For instance, additional data could be collected on whether students verified in the CPS environment. These actions were stored in a computer log file, which may provide additional information for tracking students' efforts in solving the CPS units.

Technical advances in computer-based learning systems have made greater efficiency possible by capturing more information about the problem-solving process. Finer-grained information from response time and actions were also added into CPS measurement in recent studies (e.g.,

de Boeck & Scalise, 2020; Han et al., 2023; Qiao et al., 2023). Further, many studies (e.g., von Davier et al., 2019; Han et al., 2019; Gao et al., 2022; He et al., 2021, 2023b; Ulitzsch et al., 2021) showed that process data are more appropriate to describe respondents' behaviors and strategies in interactive tasks. For example, Xiao et al. (2021) applied hidden Markov models on time-stamped action sequence data to identify the latent states and transitions between states underlying the problem-solving process. Ulitzsch et al. (2023) explored the early predict-ability of behavioral outcomes on interactive tasks with early-window clickstream data. They applied extreme gradient boosting to dynamically classify respondents who have a high probability of being out of track when solving a problem-solving task. He et al. (2023a) developed dynamic time warping method to cluster students' dynamic navigation patterns. These methods are worth further exploration to investigate the associations between sequences of actions and CPS skills and to extract sequence patterns for different CPS proficiency levels.

Considering the complexity of human-to-human interaction in collaborative conversations across countries and languages, PISA 2015 adopted the human-agent module in CPS domain. This new item type also brings challenges in test translation and fairness across countries in diversified cultural environments. It would be interesting to check for test fairness across different language groups and investigate the effect of languages in the CPS measures in a future study. The advances in text-based generative artificial intelligence applied in large language model (LLM; OpenAI, 2023) shed lights on alternative approaches to handle conversation-based assessment in the near future, which might be self-trained on different languages.

In conclusion, PISA 2015 CPS competency is a conjoint dimension of collaboration skills, functioning as a leading strand, and problem-solving skills, functioning as an essential perspective. The effectiveness of CPS depends on the ability of group members to collaborate and prioritize the success of the group over that of the individual. At the same time, this ability is a trait in each of the individual members of the group (OECD, 2013). The methods in PISA 2015 introduced in this study for collaborative item design could be applied to other collaborative human-agent items in similar settings and also challenge other researchers to refine the methodology and add extra information or data sources to get a better CPS scale. For future studies, we recommend using multivariate statistical analyses to address different aspects of CPS units and combining these analyses with process data from log files to track the process of students' learning and collaborative activities.

Acknowledgments

The author thanks Mary Louise Lennon, Henry Chen, Matthias von Davier and Hyo Jeon Shin for helpful suggestions at the initial stage of this study; the Center for Global Assessment at Educational Testing Service (ETS) for their support. All views expressed in this paper are solely those of the author and do not necessarily reflect those of the OECD or ETS.

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

Qiwei He https://orcid.org/0000-0001-8942-2047

REFERENCES

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce*. http://www.conference-board.org/pdf_free/BED-06-Workforce.pdf

Cooke, N.J., Duchon, A., Gorman, J.C., Keyton, J., & Miller, A. (2012). Preface to the special section on methods for the analysis of communication. *Human Factors: Journal of the Human Factors and Ergonomics Society, 54*, 485–488.

de Boeck, P., & Scalise, K. (2019). Collaborative problem solving: Processing actions, time, and performance. *Frontiers in Psychology, 10*, 1280.

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, *323*(5910), 66-69.

Dede, C. (2012). *Interweaving assessments into immersive authentic simulations: Design strategies for diagnostic and instructional insights*. Paper presented at the Invitational Research Symposium on Technology Enhanced Assessments. http://www.k12center.org/rsc/pdf/session4-dede-paper-tea2012.pdf

Gao, Y., Cui, Y., Bulut, O., Zhai, X., & Chen, F. (2022). Examining adults' web navigation patterns in multi-layered hypertext environments. *Computers in Human Behavior, 129*, 107142.

Greiff, S., Wüstenberg, S., Csapo, B., Demetriou, A., Hautamäki, J., Graesser, A. C., & Martin, R. (2014). Domain-general problem-solving skills and education in the 21st century. *Educational Research Review, 13*, 74-83.

Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills*. Springer.

Han, Z., He, Q., & von Davier, M. (2019). Predictive feature generation and selection using process data from PISA interactive problem-solving items: An application of random forests. *Frontiers in Psychology, 10*, 1421.

Han, A., Krieger, F., Borgonovi, F., & Greiff, S. (2023). Behavioral patterns in collaborative problem solving: a latent profile analysis based on response times and actions in PISA 2015. *Large-scale Assessments in Education*, *11*(1), 35.

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Identifying generalized behavioral patterns with sequence mining. *Computers & Education*, *166*, 104170.

He, Q., Borgonovi, F., Suárez-Álvarez, J. (2023a). Clustering Sequential Navigation Patterns in Multiple-Source Reading Tasks with Dynamic Time Warping Method. *Journal of Computer-Assisted Learning*, *39*(3), 719-736.

He, Q., Shi, Q., Tighe, E. (2023b). Predicting problem-solving proficiency with hierarchical supervised models on response process. *Psychological Test and Assessment Modeling*, *65*(1), 145-178.

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 749-776). Information Science Reference.

He, Q., von Davier, M., Greiff, S., Steinhauer, E.W., & Borysewicz, P.B. (2017). Collaborative problem-solving measures in the Programme for International Student Assessment (PISA). In A.A. von Davier, M. Zhu, & P.C. Kyllonen, (Eds.), *Innovative assessment of collaboration* (pp. 95-111). Springer.

Hirschberg, D.S. (1975). A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, *18*, 341-343.

Hirschberg, D.S. (1977). Algorithms for the longest common subsequence problem. *Journal of the ACM*, *24*(4), 664-675.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

National Academies. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_070895.pdf

National Center for Education Statistics (2015). *The nation's report card: 2015 mathematics and reading assessments.* Publication No. NCES 2015136. Washington, DC: Author.

OpenAI. (2023). ChatGPT (May 24 version) [Large language model]. https://chat.openai.com/chat/

Organisation for Economic Co-operation and Development (2013). *PISA 2015: Draft collaborative problem solving framework*. Paris, France: Author.

Organisation for Economic Co-operation and Development (2015a). *PISA 2015 released field trial cognitive items.* Paris, France: Author.

Organisation for Economic Co-operation and Development (2015b). *PISA 2015 field trial analysis report: Outcomes of the cognitive assessment (JT03371930).* Paris, France: Author.

Organisation for Economic Co-operation and Development (2017). *PISA 2015 Results (Volume V): Collaborative Problem Solving*, PISA, OECD Publishing, Paris, France.

Qiao, X., Jiao, H. & He, Q. (2023). Multiple-Group Joint Modeling of Item Responses, Response Times, and Action Counts with the Conway-Maxwell-Poisson Distribution. *Journal of Educational Measurement, 60*(2), 255-281.

Quellmalz, E.S., Timms, M.J., & Schneider, S.A. (2009). *Assessment of student learning in science simulations and games*. Paper prepared for the National Research Council Workshop on Gaming and Simulations, Washington, DC.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rosenbaum, P.R. (1988). Item bundles. *Psychometrika*, *53*(3), 349-359.

Stecher, B.M., & Hamilton, L.S. (2014). *Measuring hard-to-measure student competencies: A research and development plan,* Research Report. RAND Corporation.

Ulitzsch, E., He, Q., Ulitzsch, V., Nichterlein, A., Molter, H., Niedermeier, R., & Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *Psychometrika, 86*, 190-214.

Ulitzsch, E., Ulitzsch, V., He, Q., & Lüdtke, O. (2023). A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks. *Behavior Research Methods*, *55*, 1392–1412.

von Davier, M., Khorramdel, L., He, Q., Shin, H., & Chen, H. (2019). Developments in psychometric population models for data from innovative items. *Journal of Educational and Behavioral Statistics, 44*(6), 671-705.

Wildman, J.L., Thayer, A.L., Pavlas, D., Salas, E., Stewart, J.E., & Howse, W. (2012). Team knowledge research: Emerging trends and critical needs. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 54*, 84-111.

Wilson, M., & Adams, R.J. (1995). Rasch models for item bundles. *Psychometrika*, *60*(2), 181-198.

Xiao, Y., He, Q., Veldkamp, B.P., & Liu, H. (2021). Exploring Latent States of Problem-Solving Competence Using Hidden Markov Modeling on Process Data. *Journal of Computer-Assisted Learning, 37*(5), 1232-1247.

International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 208–223
https://doi.org/10.21449/ijate.1369290
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Evolving landscape of artificial intelligence (AI) and assessment in education: A bibliometric analysis

Nazli Ruya Taskin Bedizel [1*]

[1]Balıkesir University, Necatibey Faculty of Education, Balıkesir Türkiye

ARTICLE HISTORY

Received: Sep. 30, 2023
Accepted: Dec. 21, 2023

Keywords:
Artificial intelligence (ai),
Assessment,
Bibliometric analysis,
Education.

Abstract: The rapid evolution of digital technologies and computer sciences is ushering society into a technologically driven future where machines continually advance to meet human needs and enhance their own intelligence. Among these groundbreaking innovations, Artificial Intelligence (AI) is a cornerstone technology with far-reaching implications. This study undertakes a bibliometric review to investigate contemporary AI and assessment topics in education, aiming to delineate its evolving scope. The Web of Science Databases provided the articles for analysis, spanning from 1994 to September 2023. The study seeks to address research questions about prominent publication years, authors, countries, universities, journals, citation topics, and highly cited articles. The study's findings illuminate the dynamic nature of AI in educational assessment research, with AI firmly establishing itself as a vital component of education. The study underscores global collaboration, anticipates emerging technologies, and highlights pedagogical implications. Prominent trends emphasize machine learning, Chat GPT, and their application in higher education and medical education, affirming AI's transformative potential. Nevertheless, it is essential to acknowledge the limitations of this study, including data currency and the evolving nature of AI in education. Nonetheless, AI applications are poised to remain a prominent concern in educational technology for the foreseeable future, promising innovative solutions and insights.

1. INTRODUCTION

Progressive developments in digital technologies and computer sciences are ushering us into a future characterized by a technologically driven society, where machines are continually engineered to fulfill human requirements while also enhancing their own intelligence. Artificial Intelligence (AI) is regarded as one of the most valuable technologies, standing shoulder to shoulder with other groundbreaking innovations like robotics, virtual reality, 3D printing, and advanced networking (Chai et al., 2020; Janpla & Piriyasurawong, 2020; Kuleto et al., 2021). Technological advancements are not limited to specific regions; therefore, it is necessary to emphasize the understanding and utilization of artificial intelligence on a global scale (Bærøe et al., 2020; Grüning, 2022). Developing a collective understanding of the potential of artificial intelligence in education is crucial for ensuring equitable access to innovative educational practices worldwide (Alam et al., 2022; Bozkurt, 2023; Bozkurt et al., 2023).

*CONTACT: Nazli Ruya TASKIN BEDIZEL nazliruya@balikesir.edu.tr Balıkesir University, Necatibey Faculty of Education, Balıkesir Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

Advancing from machine learning (ML) to deep learning and ultimately to applied AI (Hassanien et al., 2020), artificial intelligence (AI) refers to the emulation of human cognitive processes, including tasks such as language translation, speech recognition, visual perception, and virtual decision-making, performed by robots and machines (Braiki et al., 2020). These cutting-edge technologies play a pivotal role in reshaping the methods and capabilities of assessment, introducing more sophisticated and nuanced approaches that align with the dynamic nature of the educational landscape (Gardner et al., 2021; Qu et al., 2022; Zehner & Hahnel, 2023). For example, by automatically creating assessments, evaluating students' written constructed responses or essays, and offering guidance and educational materials, natural language processing systems such as ChatGPT can enhance the effectiveness and efficiency of science education (Zhai, 2023).

The motivation to employ Machine Learning (ML) in scientific assessment research received a considerable boost from the National Research Council (NRC) K-12 Framework (NRC, 2012) and the Next Generation Science Standards (NGSS, 2013). Since then, there has been a strong and enthusiastic focus on the utilization of AI in educational applications (Qu et al., 2022; Toumi et al., 2018; Zhai et al., 2021). Qu et al. (2022) point out that in education, artificial intelligence encompasses various facets, including guiding learning, evaluating teaching, and refining instructional techniques, among others. Its ultimate goal is to foster teaching innovation, enrich the learning experience, and facilitate personalized education. In the realm of practical applications, AI technologies have demonstrated their efficacy beyond theoretical discussions, particularly in formative and summative assessment scenarios (Quyang et al., 2023). For example, Saito and Watanobe (2020) introduced a learning path recommendation system employing natural language processing (NLP) to assess students' programming learning performance. In addition, Erickson et al. (2020) deployed an NLP-enabled automated assessment system in a mathematics curriculum, demonstrating the capacity of AI to assess students' learning performance. Naismith et al. (2023) attempted to assess the effectiveness of using GPT-4 in evaluating the coherence of written discourse within test-taker responses on a high-stakes English proficiency test. The study revealed that GPT-4 exhibited a notable degree of accuracy in appraising the coherence of writing samples, closely matching human ratings acknowledged as the gold standard, regardless of the particular order of the prompt.

It is possible to say that the fundamental idea behind artificial intelligence (AI) in both summative and formative scenarios revolves around the concept of “machine learning." In this process, computers are essentially educated on how to discern patterns in data and are trained to execute predetermined actions based on these interpretations (Gardner et al., 2021; Zhai et al., 2021). Figure 1 presents the relationship between the intelligent assessment process and technology (Qu et al., 2022).

Figure 1. *The relationship between intelligent assessment process and technology (Qu et al., 2022, p.586)*

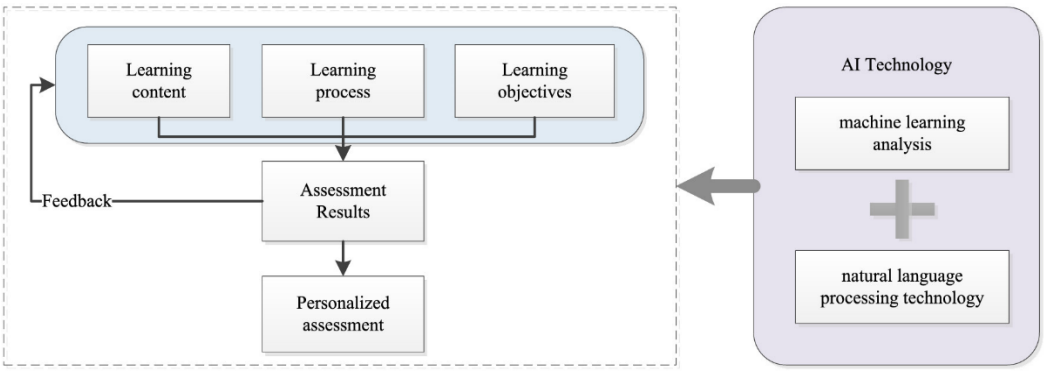


Figure 1 depicts that the advancement of intelligent assessment should be driven by the aspiration for personalized learning. It should be guided by educational theories, bolstered by machine learning analysis, and harnessed with natural language processing technology. The overarching aim is to encourage students to attain their educational objectives.

There also has been a growing debate on whether "artificial intelligence in educational assessment is a breakthrough or a buncombe and a ballyhoo?" (Gardner et al., 2021, p.1207). Zhai et al. (2020) indicate that evaluating three-dimensional learning necessitates a rethinking of assessment methodologies due to the language- and diagram-intensive characteristics of assessments grounded in scientific practices such as argumentation, explanation, and modeling. Besides, Zhai et al. (2021) put forth the argument that machine learning (ML) has the potential to enhance educational assessment by effectively capturing complex constructs, deriving precise inferences from intricate data, and simplifying the task of human grading. In parallel, commentaries and position papers (Kubsch et al., 2022; Li et al.2023; Zhai & Nehm, 2023) have extensively deliberated on the argument presented by Zhai et al. (2021). These discussions have centered around the crucial topics of equity and bias concerns, shedding light on the ethical considerations surrounding the utilization of AI in formative assessment. This issue has garnered significant attention, raising important questions about both the feasibility and desirability of incorporating AI into assessment practices. González-Calatayud et al. (2021) highlight that the field of education stands out as one of the most pertinent and pioneering areas for applying AI innovations and that the research on AI and formative assessment is essential not only for its relevance in education but also for its broader implications in shaping the future of our society.

A fundamental approach to conceptualizing any academic discipline involves a systematic examination of the associated scholarly output, as each field periodically reassesses its contributions (Agarwal et al., 2016). Studies that adeptly chart the current terrain and prevalent research directions serve as pivotal reference points for future scholarly undertakings in the discipline (Okagbue et al., 2022). Therefore, considering the growing interest and debates of utilizing AI in educational assessment practices the principal aim of the present research is to thoroughly investigate contemporary topics in AI and assessment in education with a bibliometric review, aiming to delineate its evolving scope. To reach the aims, the articles in the Web of Science Databases were examined, analyzing the articles and the emerging trends in research articles published between 1994 and September 2023. This study examines pertinent data from prior research to address the research questions outlined in Table 1.

Table 1. *Research Questions of the Study.*

	Research Question	Objective	Motivation
RQ1	Which publication years, authors, countries, universities, journals, and citation topics stand out in the field of AI and education assessment literature, and which articles have garnered the highest number of citations?	To determine the sources and authors with the highest productivity	To enhance comprehension of the leadership dynamics in the intersection of AI and educational assessment within the scientific community
RQ2	What do the bibliographic maps, graphs, and tables reveal about the data? How do they shed light on the conceptual, intellectual, and social frameworks that underpin the knowledge base necessary to advance AI in educational assessment?	To conduct a thorough analysis and present the findings concisely	To aid in grasping the current state of AI research in the field of educational assessment

2. METHOD

The present research aims to thoroughly investigate contemporary topics in AI and assessment in education, aiming to delineate its evolving scope. Numerous methods are available to analyze research trends within a field, including literature review, content analysis, meta-analysis, and meta-synthesis, among others (Kaya, 2023). The present study utilizes bibliometric analysis as a widely used and robust approach for the examination and evaluation of extensive sets of research studies conducted in a field (Zupic & Cater, 2015; Donthu et al., 2021). Bibliometric analysis allows researchers to quantitatively analyze scholarly output, such as publications, citations, and collaborations, to gain insights into the research landscape of a specific field (Agarwal et al., 2016; Donthu et al., 2021). By employing bibliometric analysis, researchers can identify interconnections, key trends, influential authors, and important research topics within a given discipline (Zupic & Cater, 2015; Okagbue et al., 2022).

2.1. Data Collection

In the present research, a chosen dataset is subjected to a quantitative examination, incorporating a bibliometric analysis. In the realm of bibliometric analysis, two primary approaches namely performance analysis and scientific mapping are commonly employed for constructing a dataset (Donthu et al., 2021). The first approach entails the selection of one or more journals, encompassing all the studies published within these journals, or including studies identified through thorough examination in the analysis. On the other hand, the second approach provides a visual representation of the interrelationships between disciplines, fields, specialties, individual papers, and authors (Small, 1999). This method is often used in studies that concentrate on specific subject areas (Donthu et al., 2021; Zupic & Cater, 2015).

In the present study, a performance analysis and scientific mapping were conducted. Performance analysis involved the utilization of carefully chosen keywords and phrases to identify relevant research. A four-step methodology, comprising keyword selection, data cleaning and formatting, preliminary analysis, and comprehensive data analysis followed in the study (Fahimnia et al., 2015). The selection process commenced with a search using keywords related to "assessment" and "AI" within the WoS Core Collection, as outlined in Table 2. The combination of "artificial intelligence" AND "assessment" ensures that articles included in the study specifically address the intersection of AI and assessment in education. This conjunction emphasizes the need for relevance to both AI and assessment topics simultaneously. The inclusion of "assess*" provides flexibility, allowing the search to capture a variety of articles that may use different forms of the term "assessment." This helps account for potential variations in terminology used across the literature. The decision to utilize the WoS Core Collection was driven by several factors (Durán-Sánchez et al.,2019). First, it is renowned for its high-quality indexes. Second, it boasts extensive coverage over a substantial timeframe. Lastly, it offers the capability to download a significant number of stored references simultaneously. To further refine the search, the research area of "Educational Education Research", “Education Scientific Disciplines” and “Psychology Educational” were applied as Web of Science Categories. Furthermore, it's important to note that only articles written in the English language were considered among the selected articles. In the data cleaning and formatting step, full records of the results were exported as an Excel file and duplications and mispresented (such as conference papers) records were removed from the list. Ultimately, 436 records were narrowed down for a more thorough examination as in Table 2.

Table 2. *Study Selection Criteria.*

Criteria	Value
1. Data Source	Web of Science Core Collection
2. Search Query	"artificial intelligence" AND "assessment" OR "assess*" (All Fields)
3. Number of Results	91270
4. Filters	Article or Review Article or Early Access (Document Types) and Education Educational Research or Education Special or Psychology Educational or Education Scientific Disciplines (Web of Science Categories) and English (Languages) and English (Languages)
5. Number of Selected Articles	436

Following the refinement of the dataset to 436 articles, an in-depth analysis of publications was conducted using the "analyze results" feature on the Web of Science platform. The examination encompassed parameters such as year of publication, country of origin, authorship, affiliations, journals, and micro-level citation topics.

Various approaches emerged for examining bibliographic data sourced from databases, including methods like citation analysis, co-author analysis, co-citation analysis, and co-word analysis (Gülmez et al., 2021).

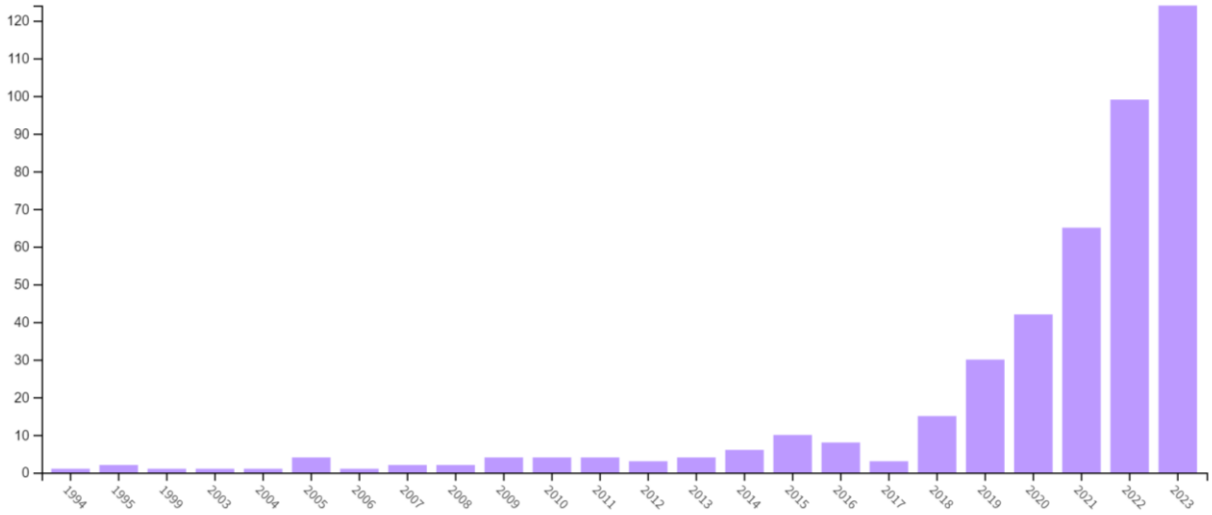
For the scientific mapping step, the maps were created to gain insights into the research topics and the various structures in the dataset (Cobo et al., 2011). Vos Viewer is used to create the co-occurrence of the keywords maps and to identify the clusters within the topic of the study. In the process of scientific mapping using VOS Viewer, various threshold values were tested to assess their influence on the outcomes. Ultimately, a minimum occurrence threshold of 5 was set to focus on significant contributions and core themes, reducing irrelevancy, enhancing interpretability, ensuring robustness, and balancing specificity and generality. This process is designed to pinpoint high-impact studies and prominent authors, as well as to scrutinize research themes that offer valuable insights for future investigations in the field.

3. RESULTS

3.1. Performance Analysis

Figure 2 provides a visual representation of the yearly distribution of articles within the chosen dataset. The trend in article productivity over the analyzed period exhibits a noticeable increase especially from 2018 to 2023. Significantly, 2019 marked a noteworthy turning point, witnessing a doubling of publications, with the release of 30 articles, thus establishing a substantial body of work. Subsequent years have consistently maintained this level of productivity, surpassing the initial threshold of 30 articles per year. It is noteworthy that more than %50 of the articles were published in 2022- 2023 and that since the year 2023 has not yet concluded, the final numbers are anticipated to exceed this current count.

Figure 2. *Distribution of publications related to AI and assessment by year*.



In Figure 3, the distribution of papers published by different countries is presented. The United States has been the most prolific country in addressing the topics of the study. Specifically, over a quarter of the articles originate from the USA. Additionally, noteworthy contributions come from countries such as the People's Republic of China, England, Spain, and Australia. This data underscores the global collaboration and collective involvement in advancing the field of AI utilization in educational assessment.

Figure 3. *Distribution of publications related to AI and assessment by country*.

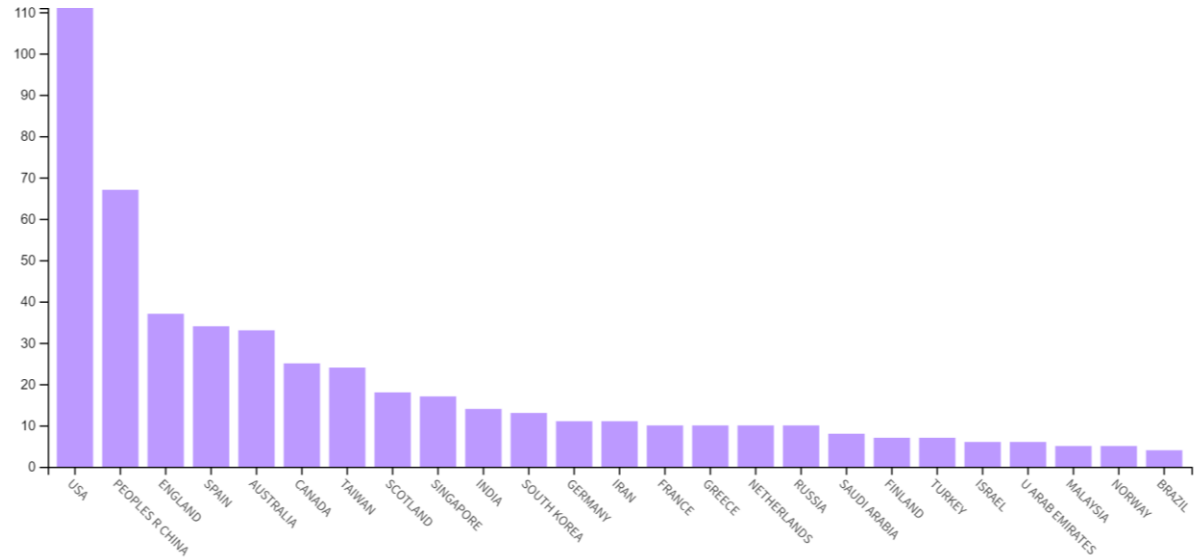


Figure 4 and Figure 5 provide insights into the authors and institutions with the highest productivity in contributing to these journals. Figure 4 reveals that the most prolific author was A.C. Graesser with 11 articles, closely followed by Z.L. Pi and J.M. Yang with 10 articles. However, it is noteworthy that 21 researchers had 9 articles each, equally contributing to the field. In Figure 5, we can observe that the institution with the highest productivity was Central China Normal University, followed by the University System of Georgia and Harvard University. It's worth noting that universities in China and the USA appear to dominate the contributions in terms of the country of origin.

Figure 4. *Distribution of publications related to AI and assessment by authors.*

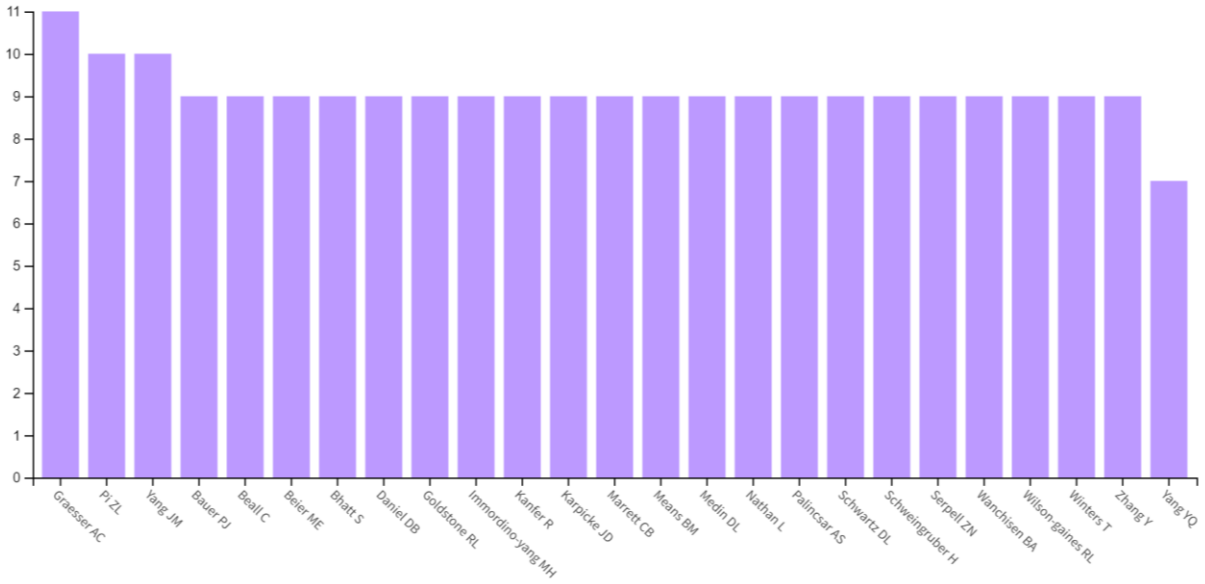


Figure 5. *Distribution of publications related to AI and assessment by affiliations.*

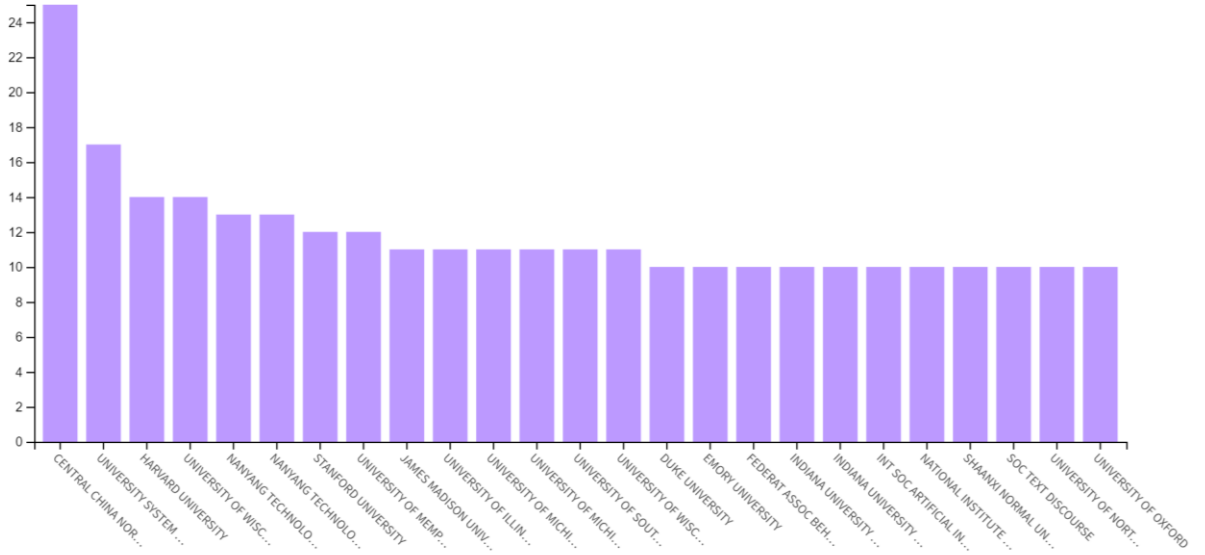


Figure 6 provides a visual representation of the journals that have published the selected articles. The figure indicates that the *Education and Information Technologies Journal and* leads with over 20 articles, followed closely by the *Education Sciences Journal and International Journal of Emerging Technologies in Learning*.

Figure 6. *Distribution of publications related to AI and assessment by publication journals.*

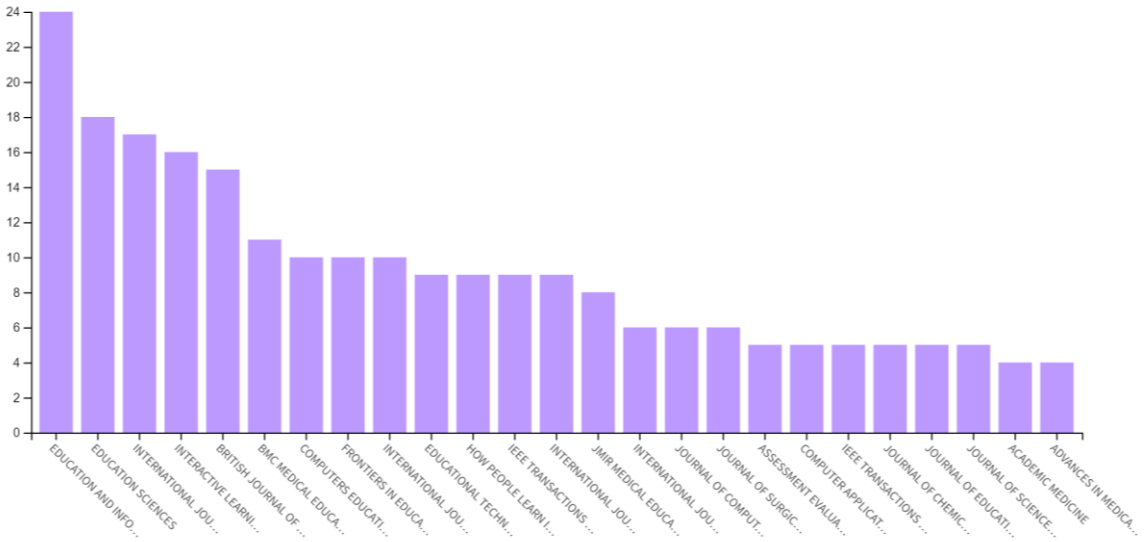
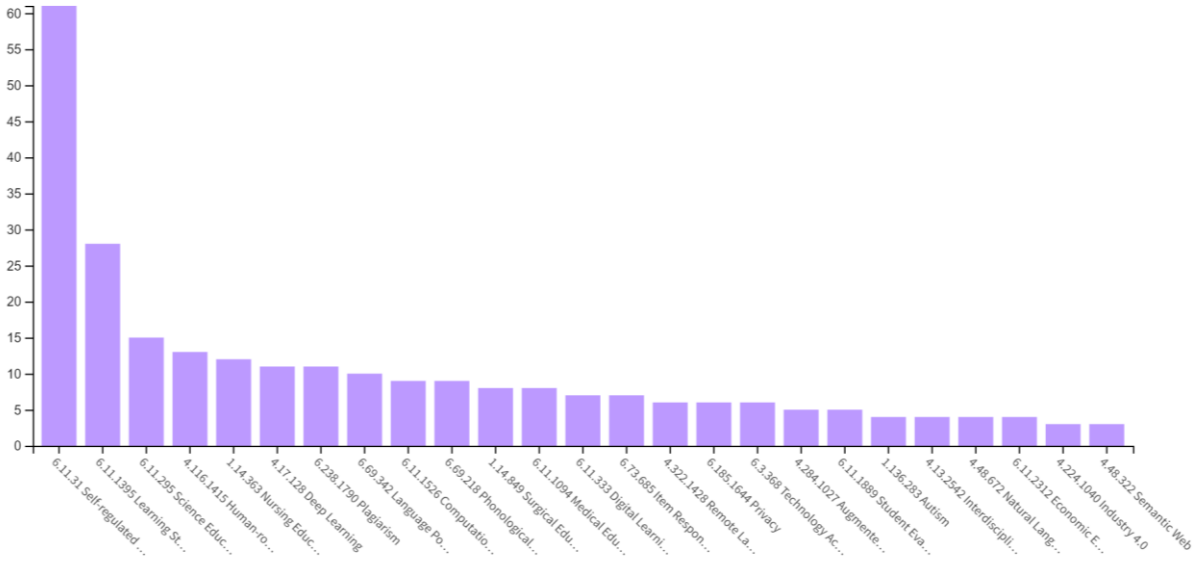


Figure 7 showcases the distribution of publications on AI and assessment, categorized by citation topics that encompass groups of related papers connected through citations. In this study, micro-topics were employed, utilizing an algorithmic tool to label each category based on the most prominent keyword. The figure reveals that the most frequently occurring citation topics revolve around *self-regulated learning*, followed by *learning styles* and *science education*. From the data, it can be inferred that the articles in the field of educational assessment and AI have had a substantial impact, particularly on research studies utilizing these keywords.

Figure 7. *Distribution of publications related to AI and assessment by citation topics (micro).*



3.2. Scientific Mapping

As noted in the methods section, scientific network maps allow for the exploration of relevant terms, research trends, and interrelationships among various concepts. These networks facilitate the detection of emerging patterns in research and the identification of areas where further investigation is needed. Figure 8 displays the keywords employed by the articles within the selected dataset, with a minimum occurrence threshold set at 5, and out of 1363 keywords, 39

meet the threshold. One prominent observation in the figure is that frequently used keywords appear larger compared to less frequently used ones.

Figure 8. *Co-occurrence of keywords in the selected dataset.*

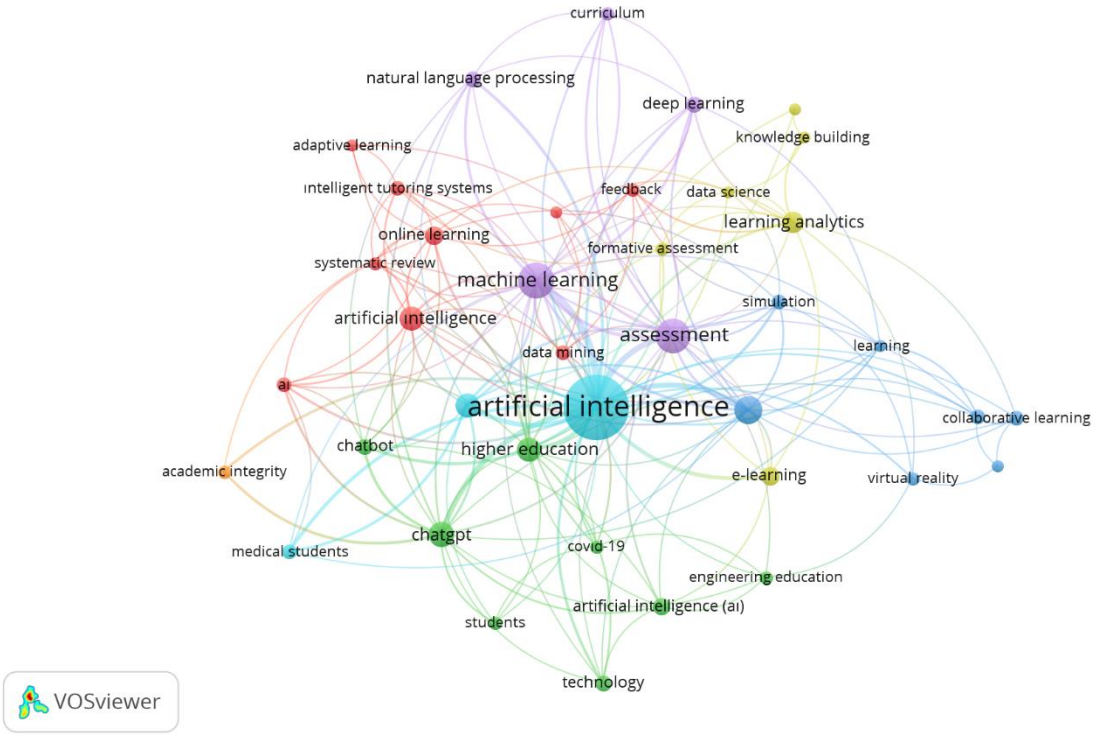
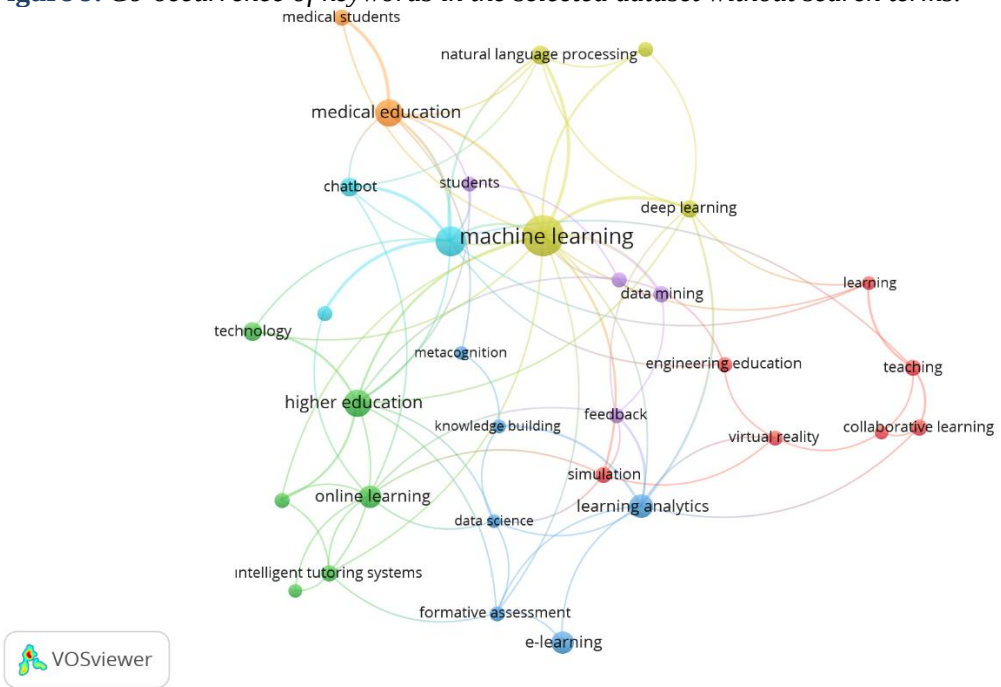


Figure 9 shows the map when artificial intelligence, AI, assessment, and education keywords were excluded. Figure 9 and Table 3 reveal the presence of seven main clusters, denoted by colors (red, green, blue, khaki, purple, turquoise, and orange), indicating interrelated words. The occurrence of these associated words and concepts within these clusters is further detailed in Table 3.

Figure 9. *Co-occurrence of keywords in the selected dataset without search terms.*



The terms that prominently feature in the analyzed papers are as follows: "machine learning" with a frequency of 40 occurrences, followed by "ChatGPT" (f=21), "higher education" (f=19), "medical education" (f=18), “online learning” (12) and e-learning (12). It can be inferred from the map that the breakdown and scientific production trends of artificial intelligence in educational assessment focused on machine learning, ChatGPT, higher education, medical education, online learning, and e-learning.

Table 3. *Clusters and co-occurrence of the keywords.*

Clusters	Co-occurrence of keywords (*f*)
Cluster 1 (7 items) (red)	Collaborative learning (7), Improving classroom Teaching (5), Learning (7), Simulation (7), Teaching (7), Virtual reality (6), engineering education (6)
Cluster 2 (6 items) (green)	Higher education (19), Online learning (12), Technology (9), Intelligent tutoring Systems (7), Systematic review (6), Adaptive learning (5)
Cluster 3 (6 items) (blue)	Learning analytics (14), E-Learning (12), Formative assessment (6), Data science (5), Knowledge building (5), Metacognition (5)
Cluster 4 (4 items) (khaki)	Machine learning (40), Deep learning (8), Natural language Processing (9), Curriculum (6)
Cluster 5 (4 items) (purple)	Data mining (7), Feedback (6), Covid-19 (6), Students (6)
Cluster 6 (3 items) (turquoise)	Chatgpt (21), Chatbot (9), Academic integrity (6)
Cluster 7 (2 items) (orange)	Medical education (18), Medical students (7)

As can be inferred from Table 3 **Cluster 1** revolves around the concept of collaborative learning, virtual reality, and improving classroom teaching. It suggests that collaborative and immersive learning experiences are integral to AI in educational assessment. The inclusion of keywords like simulation and engineering education indicates a focus on practical and hands-on learning experiences (e.g. Winkler-Schwarz et al., 2019). The emphasis on teaching and learning in this cluster suggests a commitment to enhancing the educational experience through AI-driven methods.

Cluster 2 centers on higher education and online learning, emphasizing the importance of AI in these contexts. It includes terms like technology, intelligent tutoring systems, and systematic review, highlighting a scholarly approach to incorporating AI into higher education (Sharma & Harkishan, 2022; Zawacki-Richter et al., 2019). The cluster's focus on adaptive learning underscores the desire to tailor education to individual student needs (Sharma et al., 2019).

Cluster 3 focuses on learning analytics, e-learning, and formative assessment, indicating a strong emphasis on data-driven educational practices. Keywords like data science and metacognition suggest a rigorous analytical approach to educational assessment (Wood et al., 2021). The presence of terms like knowledge building reflects a community dedicated to advancing pedagogy through AI and data.

Cluster 4 centers on machine learning, deep learning, and natural language processing are foundational to this cluster, highlighting the centrality of advanced AI techniques in educational assessment. The curriculum is a critical keyword, indicating the integration of AI into educational curricula. The prevalence of machine learning-related terms suggests a community of researchers and practitioners focused on AI's potential in education.

Cluster 5 includes data mining, feedback, and mentions of COVID-19, highlighting the importance of data-driven decision-making and adaptability in the face of challenges (Yang et al., 2023). The presence of keywords related to students suggests a student-centered approach to AI in education. The focus on feedback indicates a concern for enhancing the learning

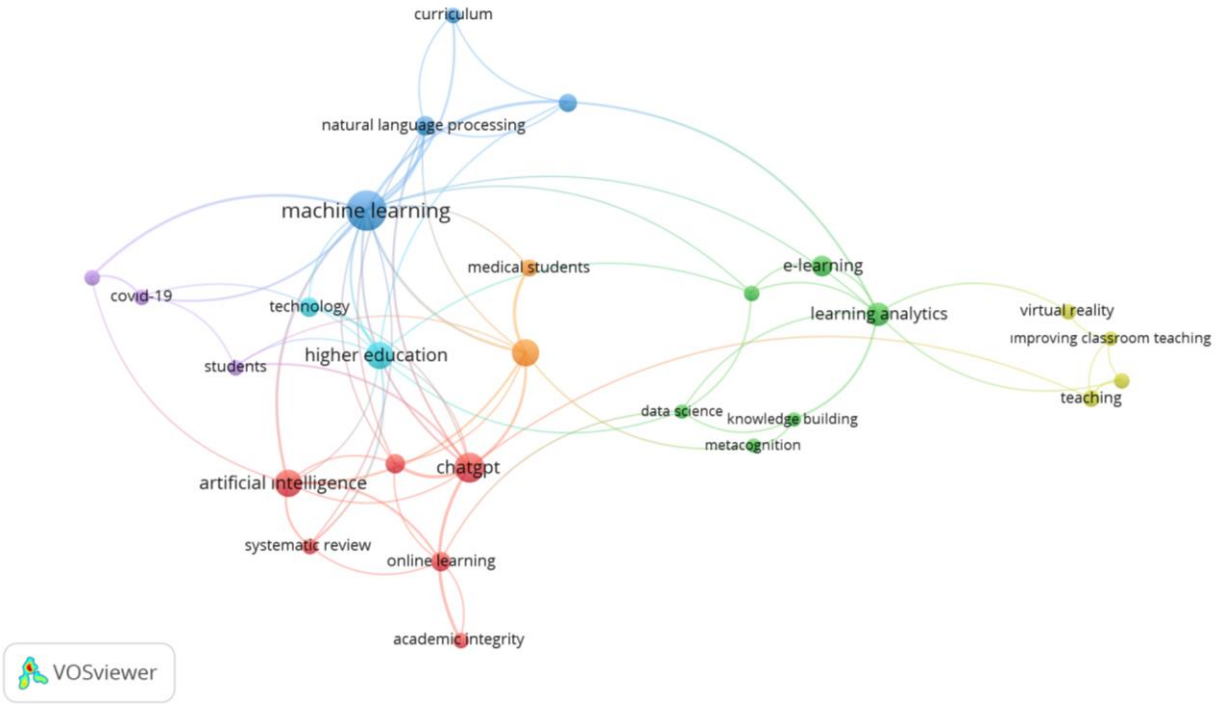
experience through assessment and improvement.

Cluster 6 focuses on ChatGPT and chatbots, emphasizing the role of conversational AI in educational assessment. Academic integrity is a key term, suggesting a focus on ethical considerations in AI-driven assessment (Lancaster, 2023). The prominence of chatbot-related keywords implies the existence of communities exploring AI-driven chat systems in education.

Cluster 7 includes medical education and medical students as the core themes, highlighting the application of AI in the medical field. This cluster reflects a specialized area of research (e.g., Civaner et al., 2022; Tolsgaard et al., 2023; Winkler- Schwarz et al., 2019) within AI in education, focusing on medical training. The emphasis on medical education suggests a dedicated community of researchers and educators in this domain.

Since the field showed a breakdown in 2018, the articles from the beginning of 2018 until September 2023 were also examined as a network map in Vos Viewer. Figure 10 presents this map. The depicted figure highlights the dominance of certain keywords such as "machine learning," "ChatGPT," "higher education," "medical education," and "learning analytics" within the field of artificial intelligence in educational assessment.

Figure 10. *Co-occurrence of keywords between 2018 and 2023.*



4. DISCUSSION and CONCLUSION

The present article tried to thoroughly investigate contemporary topics in AI and assessment in education with a bibliometric review, aiming to delineate its evolving scope. In conclusion, this study's findings have illuminated the remarkable growth and global collaboration within the field of artificial intelligence in educational assessment in recent years. The surge in publications, the prominence of specific keywords, and the interconnected clusters of terms collectively underscore the dynamic and evolving nature of research in this domain. As highlighted by Latif et al. (2023), the study affirms that Artificial Intelligence (AI) has firmly established itself as an integral element of educational practice and assessment. This evolving landscape suggests that educators and researchers should continuously adapt to the changing educational technology environment to harness the potential of AI effectively.

The findings of the study underscored the significance of global collaboration, with

contributions from various countries and institutions. As the field continues to evolve, likely, emerging technologies and innovative approaches will likely further shape the landscape of AI in educational assessment, providing valuable insights and tools for educators and researchers alike. A vast amount of research on the field (e.g., Baker & Yacef, 2009; Siemens & Baker, 2012; Baker & Inventado, 2014) covers various aspects of AI in education, including design-based research, learning analytics, cognitive tutors, stealth assessment, and ethical considerations. They also highlight the contributions from different countries and institutions, emphasizing the collaborative nature of the field. This collaborative spirit can lead to more comprehensive and effective AI applications in education.

Prominent trends identified in the study encompass a concentrated emphasis on machine learning, ChatGPT, and their application in higher education and medical education. This reflects a concerted endeavor to harness AI's capabilities within these specific domains. Reinforcing these observations, Zawacki-Richter et al. (2019) conducted a systematic review that delved into the research on artificial intelligence applications in higher education. Their findings underscore the potential transformative impact of AI on higher education institutions. Moreover, they shed light on the substantial investments and keen interest in AI from both private companies and public-private partnerships. This corroborates the study's assertion that AI's influence in higher education remains a significant focal point, further emphasizing the importance of AI in this sector. Sapci & Sapci (2020) also contributed to the understanding of AI in education, particularly in the context of medical and health informatics. The systematic review explores the integration of AI training into medical and health informatics curricula, indicating a growing recognition of the importance of AI education in these fields. In addition, Bozkurt et al. (2021) provided a comprehensive review of AI studies in education over the past half-century. The authors used a systematic review approach and employed social network analysis and text-mining approaches to identify key research clusters and themes. The study identified three research clusters, one of which is focused on artificial intelligence. Within this cluster, the study highlights the theme of adaptive learning and personalization of education through AI-based practices, which aligns with the present study. Educators and researchers should stay informed about these developments to leverage the latest tools and insights for improved teaching and assessment.

The prevalence of citation topics such as self-regulated learning, learning styles, and science education underscores a substantial focus on pedagogical aspects within the field. This emphasis is in line with the recognition of how learning styles can significantly influence a variety of assessment methods and practices, as discussed by Calatayud et al. (2021). Additionally, it aligns with the potential for artificial intelligence to bring about transformative changes in the delivery and evaluation of education, which holds the promise of enhancing educational outcomes for students, as articulated by Owan et al. (2023). Consequently, the integration of artificial intelligence into this educational domain is not only a logical step but also an expected and prominent development. Future research and implementations should prioritize pedagogical effectiveness.

The clusters created by Vos Viewer collectively represent the multifaceted nature of AI in educational assessment (Baker & Yacef, 2009; Siemens & Baker, 2012; Luckin et. al., 2016), with each cluster contributing to the broader knowledge base necessary to advance the field. They underscore the diverse applications of AI, from collaborative and immersive learning experiences to data-driven decision-making and personalized education. Moreover, they emphasize the importance of ethical considerations and the potential for AI to revolutionize education in various domains (Zhai & Nehm, 2023), including medicine. Understanding these clusters is crucial for researchers, educators, and policymakers seeking to leverage AI's potential in educational assessment effectively.

Regarding the study's limitations, it should be noted that the research relies on data available up to a specific point in time. Since then, new publications and emerging trends may have surfaced, potentially escaping the scope of this analysis. Furthermore, the study predominantly concentrates on bibliometric analysis and the tracking of keyword trends. It does not delve into the qualitative dimensions of research or provide an in-depth exploration of the specific applications of AI in education. It is also important to recognize that while the study does identify prevailing trends, it may not comprehensively capture the full spectrum of AI applications in education across diverse contexts and regions. Consequently, caution should be exercised when attempting to generalize the findings to all educational settings. Lastly, the study offers insights into potential future developments in AI in education. However, it is essential to acknowledge that the actual trajectory of AI's role in education may be subject to a multitude of unpredictable influences, including advancements in technology, alterations in policy, and shifts in societal dynamics.

In conclusion, as indicated by Zawacki-Richter et al., (2019) the complete outcomes of AI progress remain unpredictable at this time. However, it appears probable that AI applications will emerge as a prominent concern in the realm of educational technology for the next two decades. Moreover, the influence of AI within education continues to broaden and deepen, promising innovative solutions and insights for the field.

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

Nazli Ruya Taskin Bedizel https://orcid.org/0000-0001-6027-719X

REFERENCES

Agarwal, A., Durairajanayagam, D., Tatagari, S., Esteves, S., Harlev, A., Henkel, R., Roychoudhury, S., Homa, S., Puchalt, N., Ramasamy, R., Majzoub, A., Ly, K., Tvrda, E., Assidi, M., Kesari, K., Sharma, R., Banihani, S., Ko, E., Abu-Elmagd, M., … Bashiri, A. (2016). Bibliometrics: tracking research impact by selecting the appropriate metrics. *Asian Journal of Andrology, 18*(2), 296. https://doi.org/10.4103/1008-682x.171582

Alam, A., Hasan & Raza, M. (2022). Impact of artificial intelligence (AI) on education: changing paradigms and approaches. *Towards Excellence*, 281-289. https://doi.org/10.37867/te140127

Bærøe, K., Miyata-Sturm, A., & Henden, E. (2020). How to achieve trustworthy artificial intelligence for health. *Bulletin of the World Health Organization, 98*(4), 257-262. https://doi.org/10.2471/blt.19.237289

Baker, R.S., & Inventado, P.S. (2014). Educational data mining and learning analytics. *Learning Analytics*, 61-75. https://doi.org/10.1007/978-1-4614-3305-7_4

Baker, R.S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3-17.

Bozkurt, A. (2023). Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift. *Asian Journal of Distance Education, 18*(1). Retrieved from https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/718

Bozkurt, A., Karadeniz, A., Baneres, D., Guerrero-Roldán, A.E., & Rodríguez, M.E. (2021). Artificial intelligence and reflections from educational landscape: a review of AI studies in half a century. *Sustainability, 13*(2), 800. https://doi.org/10.3390/su13020800

Bozkurt, A., Xiao, J., Lambert, S., Pazurek, A., Crompton, H., Koseoglu, S., Farrow, R., Bond, M., Nerantzi, C., Honeychurch, S., Bali, M., Dron, J., Mir, K., Stewart, B., Costello, E.,

Mason, J., Stracke, C.M., Romero-Hall, E., Koutropoulos, A., … Jandrić, P. (2023). Speculative futures on ChatGPT and generative artificial intelligence (AI): A collective reflection from the educational landscape. *Asian Journal of Distance Education, 18*(1), 53-130. https://doi.org/10.5281/zenodo.7636568

Braiki, B.A., Harous, S., Zaki, N., & Alnajjar, F. (2020). Artificial intelligence in education and assessment methods. *Bulletin of Electrical Engineering and Informatics, 9*(5), 1998-2007. https://doi.org/10.11591/eei.v9i5.1984

Chai, C.S., Wang, X., & Xu, C. (2020). An extended theory of planned behavior for the modelling of Chinese secondary school students' intention to learn artificial intelligence. *Mathematics, 8*(11), 2089. https://doi.org/10.3390/math8112089

Civaner, M.M., Uncu, Y., Bulut, F., Chalil, E.G., & Tatli, A. (2022). Artificial intelligence in medical education: a cross-sectional needs assessment. *BMC Medical Education, 22*(1), 772. https://doi.org/10.1186/s12909-022-03852-3

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology, 62*(7), 1382-1402. https://doi.org/10.1002/asi.21525

Donthu, N., Kumar, S., Pandey, N., Pandey, N., & Mishra, A. (2021). Mapping the electronic word-of-mouth (eWOM) research: A systematic review and bibliometric analysis. *Journal of Business Research, 135,* 758-773. https://doi.org/10.1016/j.jbusres.2021.07.015

Durán-Sánchez, A., Del Río-Rama, M. de la C., Álvarez-García, J., & García-Vélez, D.F. (2019). Mapping of scientific coverage on education for entrepreneurship in higher education. *Journal of Enterprising Communities: People and Places in the Global Economy, 13*(1/2), 84-104. https://doi.org/10.1108/jec-10-2018-0072

Erickson, J.A., Botelho, A.F., McAteer, S., Varatharaj, A., & Heffernan, N.T. (2020). The automated grading of student open responses in mathematics. In C. Rensing, & H. Drachsler (Eds.), *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 615–624). Association for Computing Machinery. https://doi.org/10.1145/3375462.3375523

Fahimnia, B., Sarkis, J., & Davarzani, H. (2015). Green supply chain management: A review and bibliometric analysis. *International Journal of Production Economics, 162*, 101-114. https://doi.org/10.1016/j.ijpe.2015.01.003

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'. *Journal of Computer Assisted Learning, 37*(5), 1207-1216. https://doi.org/10.1111/jcal.12577

González-Calatayud, M.L., Fernández, C., & Meneses, J. (2019). Learning styles and educational assessment: A systematic review. *Frontiers in Psychology, 10*, 2381. https://doi.org/10.3389/fpsyg.2019.02381

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: a systematic review. *Applied Sciences, 11*(12), 5467. https://doi.org/10.3390/app11125467

Grüning, D.J. (2022). Synthesis of human and artificial intelligence: review of “how to stay smart in a smart world: why human intelligence still beats algorithms” by Gerd Gigerenzer. *Futures & Foresight Science, 4*(3-4). https://doi.org/10.1002/ffo2.137

Gülmez, D., Özteke, İ., & Gümüş, S. (2021). Overview of Educational Research from Turkey Published in International Journals: A Bibliometric Analysis. *Education & Science/Eğitim ve Bilim, 46*(206), 1-27. https://doi.org/10.15390/EB.2020.9317

Hassanien, A., Darwish, A., & El-Aska, H. (2020). *Machine Learning and Data Mining in Aerospace Technology*. Springer Nature Switzerland AG: Cham, Switzerland.

Janpla, S., & Piriyasurawong, P. (2018). The development of problem-based learning and concept mapping using a block-based programming model to enhance the programming competency of undergraduate students in computer science. TEM Journal, 7(4), 708.

Kaya, S. (2023). A bibliometric journey into research trends in curriculum field: Analysis of two journals. *International Journal of Assessment Tools in Education, 10*(3), 496-506. https://doi.org/10.21449/ijate.1278728

Kubsch, M., Czinczel, B., Lossjew, J., Wyrwich, T., Bednorz, D., Bernholt, S., Fiedler, D., Strauß, S., Cress, U., Drachsler, H., Neumann, K., & Rummel, N. (2022). Toward learning progression analytics — Developing learning environments for the automated analysis of learning using evidence centered design. *Frontiers in Education, 7*. https://doi.org/10.3389/feduc.2022.981910

Kuleto, V., Ilić, M., Dumangiu, M., Ranković, M., Martins, O.M.D., Păun, D., & Mihoreanu, L. (2021). Exploring opportunities and challenges of artificial intelligence and machine learning in higher education institutions. *Sustainability, 13*(18), 10424. https://doi.org/10.3390/su131810424

Lancaster, T. (2023). Artificial intelligence, text generation tools and ChatGPT–does digital watermarking offer a solution?. *International Journal for Educational Integrity, 19*(1), 10. https://doi.org/10.1007/s40979-023-00131-6

Latif, E., Mai, G., Nyaaba, M., Wu, X., Liu, N., Lu, G.,Li, S., Liu, T., & Zhai, X. (2023). Artificial general intelligence (AGI) for education. *arXiv preprint* arXiv:2304.12479. https://doi.org/10.48550/arXiv.2304.12479

Li, T., Reigh, E., He, P., & Adah Miller, E. (2023). Can we and should we use artificial intelligence for formative assessment in science? *Journal of Research in Science Teaching, 60*(6), 1385-1389. https://doi.org/10.1002/tea.21867

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L.B. (2016). Intelligence unleashed: An argument for AI in education. *Journal of Computer Assisted Learning, 32*(3), 201-210. https://doi.org/10.1111/jcal.12140

Naismith, B., Mulcaire, P., & Burstein, J. (2023, July). Automated evaluation of written discourse coherence using GPT-4. *In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 394-403).

National Research Council. (2012). A Framework for K-12 Science Education. https://doi.org/10.17226/13165

National Research Council. (2013). Next Generation Science Standards: For states, by states. https://doi.org/10.17226/18290

Okagbue, E.F., Ezeachikulo, U.P., Nwigwe, E.O., & Juma, A.A. (n.d.). Machine learning and artificial intelligence in education research: a comprehensive overview of 22 years of research indexed in the scopus database. https://doi.org/10.21203/rs.3.rs-1845778/v1

Ouyang, F., Dinh, T.A., & Xu, W. (2023). A systematic review of AI-driven educational assessment in stem education. *Journal for STEM Education Research, 6*(3), 408-426. https://doi.org/10.1007/s41979-023-00112-x

Owan, V.J., Abang, K.B., Idika, D.O., Etta, E.O., & Bassey, B.A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(8), em2307. https://doi.org/10.29333/ejmste/13428

Qu, J., Zhao, Y., & Xie, Y. (2022). Artificial intelligence leads the reform of education models. *Systems Research and Behavioral Science, 39*(3), 581-588. https://doi.org/10.1002/sres.2864

Saito, T., & Watanobe, Y. (2020). Learning path recommendation system for programming education based on neural networks. *International Journal of Distance Education Technologies (IJDET), 18*(1), 36-64. https://doi.org/10.4018/IJDET.2020010103

Sapci, A.H., & Sapci, H.A. (2020). Artificial intelligence education and tools for medical and health informatics students: systematic review. *JMIR Medical Education, 6*(1), e19285. https://doi.org/10.2196/19285

Sharma, K., Papamitsiou, Z., & Giannakos, M. (2019). Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach. *British Journal of Educational Technology, 50*(6), 3004-3031. https://doi.org/10.1111/bjet.12854

Sharma, P., & Harkishan, M. (2022). Designing an intelligent tutoring system for computer programing in the Pacific. *Education and Information Technologies, 27*(5), 6197-6209. https://doi.org/10.1007/s10639-021-10882-9

Siemens, G., & Baker, R.S.J.d. (2012). Learning analytics and educational data mining. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. https://doi.org/10.1145/2330601.2330661

Tolsgaard, M.G., Pusic, M.V., Sebok-Syer, S.S., Gin, B., Svendsen, M.B., Syer, M.D., Brydges, R., Cuddy, M.M., & Boscardin, C.K. (2023). The fundamentals of Artificial Intelligence in medical education research: AMEE Guide No. 156. *Medical Teacher, 45*(6), 565-573. https://doi.org/10.1080/0142159x.2023.2180340

Toumi, Y., Bengherbia, B., Lachenani, S., & Ould Zmirli, M. (2022). FGPA implementation of a bearing fault classification system based on an envelope analysis and artificial neural network. *Arabian Journal for Science and Engineering, 47*(11), 13955-13977. https://doi.org/10.1007/s13369-022-06599-7

Wood, E.A., Ange, B.L., & Miller, D.D. (2021). Are we ready to integrate artificial intelligence literacy into medical school curriculum: students and faculty survey. *Journal of Medical Education and Curricular Development, 8*. https://doi.org/10.1177/23821205211024078

Yang, Y., Zheng, Z., Zhu, G., & Salas-Pilco, S.Z. (2023). Analytics-supported reflective assessment for 6th graders' knowledge building and data science practices: An exploratory study. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13308

Zawacki-Richter, O., Marín, V.I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education - where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1). https://doi.org/10.1186/s41239-019-0171-0

Zehner, F., & Hahnel, C. (2023). Artificial intelligence on the advance to enhance educational assessment: Scientific clickbait or genuine gamechanger?. *Journal of Computer Assisted Learning, 39*(3), 695-702. https://doi.org/10.1111/jcal.12810

Zhai, X. (2023, August 28- September 1). *ChatGPT for next generation science learning* [Paper presentation]. The 15th Conference of the European Science Education Research Association (ESERA), Cappadocia, Türkiye.

Zhai, X., & Nehm, R.H. (2023). AI and formative assessment: The train has left the station. *Journal of Research in Science Teaching, 60*(6), 1390-1398. https://doi.org/10.1002/tea.21885

Zhai, X., Shi, L., & Nehm, R.H. (2021). A meta-analysis of machine learning-based science assessments: factors impacting machine-human score agreements. *Journal of Science Education and Technology, 30*(3), 361-379. https://doi.org/10.1007/s10956-020-09875-z

Zhai, X., Haudek, K.C., Shi, L., Nehm, R.H., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430-1459. https://doi.org/10.1002/tea.21658

Zupic, I., & Čater, T. (2015). Bibliometric methods in management and organization. *Organizational Research Methods, 18*(3), 429-472. https://doi.org/10.1177/1094428114562629

International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 224–242
https://doi.org/10.21449/ijate.1368600
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

The use of on-screen calculator as a digital tool in technology-enhanced items

Ozge Ersan[1*], **Burcu Parlak**[1]

[1]Ministry of National Education, General Directorate of Measurement, Evaluation and Examination Services, Ankara. Türkiye

ARTICLE HISTORY

Received: Sep. 29, 2023

Accepted: Dec. 18, 2023

Keywords:

Technology-enhanced items,
Calculator use,
Mathematics achievement,
Trends in mathematics and science study (TIMSS),
Problem solving and inquiry tasks.

Abstract: In this study, the effect of using on-screen calculators on eighth grade students' performance on two TIMSS 2019 Problem Solving and Inquiry Tasks items considered as examples of technology-enhanced items administered on computers was examined. For this purpose, three logistic regression models were run where the dependent variables were giving a correct response to the items and the independent variables were mathematics achievement and on-screen calculator use. The data of student from 12 countries and 4 benchmarking participants were analyzed and some comparisons were made based on the analyses. The results indicate that using on-screen calculators is positively associated with higher odds of giving correct responses for both items above and beyond students' mathematics achievement scores. The results of this study promote the inclusion of on-screen calculator as a digital tool in technology-enhanced items that require problem solving.

1. INTRODUCTION

Item types used in the assessments have evolved to be richer in technological features following the widespread use of computerized testing. These new types of items differ from the conventional multiple-choice (MC) and constructed-response (CR) items in terms of technological innovations. To define technology-enhanced (TE) items, Parshall et al. (2010) provided seven facets that each of these facets can vary at different levels of innovations in the items. These facets are: (a) assessment structure, (b) response action, (c) interactivity, (d) media inclusion, (e) fidelity, (f) complexity, and (g) scoring method.

The assessment structure describes how a TE item is formatted. A taxonomy for assessment structure for e-assessment items are described in the literature (Scalise & Gifford, 2006, p. 9). The structure of TE items can vary from the most constrained (multiple-choice) form to the least constrained (presentation/portfolio) forms, and the items in between were referred to as intermediate constraint items (selection/identification, reordering/rearrangement, substitution/correction, completion, construction). Response action indicates how item responses were collected such as by mouse clicks, keyboard typing, or voice recording. Interactivity refers to how test takers interact with the item such as running a science simulation

*CONTACT: Ozge ERSAN ozge.ersan09@gmail.com Milli Eğitim Bakanlığı, Ölçme Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü, Ankara. Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

or using item tools such as magnifier, highlighter or ruler for the item. Media inclusion indicates that a graphic, picture, short animation, or a sound clip may be added to the item stem or response options. Fidelity refers to the realistic and accurate representation of a scenario, task, graph, or picture. Complexity of an item indicates how each facet of innovations are combined during item development phase such as item structure, number of response options, number of supporting materials, multiple forms of response actions, as well as the design of item interface. Last, scoring method indicates a strategy for translating all inputs of the test taker into a quantitative score relevant to the measured construct (Parshall et al., 2010).

A special case of TE items, scenario based items (a.k.a. task based simulations, performance tasks), are integrated item set developed around a common scenario. The common scenario or each item relevant to the scenario may include a passage, a video clip, an animation, a graph, or a small simulation run by the test taker. Since scenario based items are generally developed to represent real life problems and tasks, they offer a potential for high fidelity in contributing to the validity of use and interpretation of test scores (Russell & Moncaleano, 2019; Sireci & Zenisky, 2006). Despite advantages scenario based items offer, there are also measurement challenges they may pose. Developing scenario based items is more challenging and expensive when compared to stand-alone items, as a result they tend to be fewer in item pools posing validity threat for repeated item exposure and memory effects (Bryant, 2017; Impara & Foster, 2006; Sireci & Zenisky, 2006). Furthermore, complex structure of scenario based items may require detailed consideration if partial credit scoring is required or what kind of scoring rule should be applied (Betts et al., 2022; Clyne, 2015; Lorié, 2016).

1.1. Technological Innovations in Trends in Mathematics and Science Study

TIMSS is an international assessment administered every four years starting from 1995 that measures mathematics and science achievements of fourth and eighth grade students. A transition from paper based assessment to digital assessment started in 2019 which is called eTIMSS 2019. Along with the digital transition, technological features were added to the items in the existing pool or new TE items were developed accordingly (Martin et al., 2020). Innovations in eTIMSS 2019 were also observed in new interactive item types called Problem Solving and Inquiry (PSI) tasks which were technology-enhanced scenario based items. By using PSI tasks, IEA aimed to extend the coverage and enhance the measurement of the TIMSS mathematics and science assessment frameworks benefitting from the features of computerized assessments, especially in the applying and reasoning cognitive domains. PSI tasks simulate real world and laboratory situations where students can apply and combine their content knowledge, skills, reasoning, and interpretation of a given situation by solving a mathematics problem, running a scientific experiment or running multiple steps of a simulation. PSI tasks involve visually attractive, interactive scenarios that require students follow a series of tasks or TE items with various response actions (e.g., number pad, drag and drop, graphing tools, and free drawings) in an adaptive and responsive way that would bring them toward a final solution or product (Mullis et al., 2021).

With increased fidelity in eTIMSS 2019 items and PSI tasks, a ruler and calculator were also available to the students at eighth grade as part of the on-screen interface. The on-screen calculator included the four basic functions (+, −, ×, ÷) and a square root key. Since a standardized ruler and calculator was available on the test screen, students were not allowed to bring their own rulers and calculators (Mullis & Martin, 2017).

1.2. Tool Use in Technology-Enhanced Items

Technological innovations of computerized items also include tools offered with the item such as magnifier, digital pen for highlighting or taking notes on a digital scratchpad, ruler, or a calculator. Some tools may be compulsory to use for the test taker to be able to correctly respond

to an item (Salles et al., 2020); they can also be available for all the items or test takers as universal tools across the entire test (WIDA, n.d.). Although examining how such tools contribute to the test taking experience in paper-pencil or classroom assessments has a long history of research, studies for digital tools in computerized assessments are limited.

Process data collected during the test administration now offer information regarding the students' use of tools. Process data may include information for which tool is used, frequencies of each tool using, or patterns of tool using to understand the test-taking strategies of students, to collect evidence for suspicious test-taking activities or to collect evidence for fairness issues. Analyzing process data regarding the use of digital tools can also contribute to item and test development processes as they provide clues for how to ease test-taking processes, eliminating construct-irrelevant variances and increasing the fidelity and validity of the item. For instance, Salles et al., (2020) showed that test takers who responded to a mathematical item with a graph correctly tended to use a digital pen for taking notes on the graph. Another study on computer based office simulation tests showed that successful test takers tended to use notepad and spreadsheets helping computation more efficiently (Ludwig & Rausch, 2022).

1.3. On-Screen Calculators as a Technological Innovation

A group of researchers who are against the use of calculators in mathematics classes before high school stated that calculation skills and understanding of mathematics concepts may be negatively affected by the use of calculators during learning (e.g., Dick, 1988; Hopkins, 1992; Plunkett, 1978). Another perspective of research indicates that calculators can ease the learning process as they still would require the students to improve their mental computation strategies anyway. Similarly, when the student is solving a problem and faces a complex calculation in the middle of a problem, the student's work flow does not need to be interrupted due to hand calculation (Sparrow et al., 1994; Vasquez & McCabe, 2002; Williams, 1987).

Parallel to the idea of using calculators during the teaching and the learning of mathematics, researchers debated the use of calculators during assessments as well. The National Council of Teachers of Mathematics (NCTM) states that when on-screen calculators are used appropriately, calculators can positively contribute to students' fluency in numbers, operations, and estimation skills (2015). According to early research findings conducted by Hopkins (1992), numbers in problems can be made more compatible with realistic situations, making the use of calculators more appropriate. Additionally, calculators can increase motivations in students' test taking (Ellington, 2003).

Test developers and other test score users should be aware that the frequency of calculator use may have an effect on students' performance in assessments (Tarr et al., 2000). Additionally, the availability of on-screen calculators should be determined depending on item types and complexity level of the items (Cohen & Kim, 1992; Loyd, 1991). For some item types, an on-screen calculator should not be provided depending on the construct being measured and for some items the calculators may not be needed. For instance, Walcott and Stickles (2012) conducted a study using eighth grade level NAEP data that included two types of items — problem solving items and noncomputational mathematics concept items— where they studied the effect of calculator use and item types. The results showed that students who used calculators had significantly better performance on problem solving items when compared to students who did not use calculators. On the other hand, calculators were not used by the majority of the students for noncomputational mathematics concept items and the ones used consistently performed worse on the test.

In summary, research shows that the calculator can improve students' fluency in numbers, operation and estimation skills that may contribute to the development of complex problem-solving and higher-order thinking skills as well as increase motivations in the students' test taking. Additionally, computerized assessments can control the calculator effect providing the

same on-screen calculators to all students suitable to the given item type and grade level. Yet, test developers should be aware that calculator use should not change the measured construct and therefore an on-screen calculator may only be available for specific items (Wolfe, 2010). Finally, further validity research is needed to examine the extent to which frequency of calculator use affects test scores to ensure equity across cultures or education systems and whether the on-screen calculator contributes to students' mathematics performance.

1.4. Purpose of the Study

As a digital tool, an on-screen calculator for mathematics items including PSI tasks in eTIMSS 2019 was available to the students. In PSI tasks, while there were two successive items administered that were essentially developed to be calculator neutral, calculators can also help problem solving process.

Preliminary analysis results of calculator use relevant to these items were reported in eTIMSS 2019 PSI report (Mullis et al., 2021). According to the report, around 88% and 84% of students used the on-screen calculator for the first and second items respectively among the students who answered the items correctly.

Therefore, preliminary findings imply that availability and use of calculators may be helpful for responding to TE items correctly; however, further research is needed to examine the extent to which use of on-screen calculators contributes to student responses above and beyond mathematics proficiency. If a significant contribution is observed, this finding will provide some evidence for item and test development endeavors in terms of making on-screen calculators available as part of innovations in TE items. To serve this purpose, the following research question was investigated: "To what extent does an on-screen calculator available for two TE items in eTIMSS 2019 PSI tasks explain eighth graders' probability of giving correct responses above and beyond their mathematics achievements?"

2. METHOD

2.1. Data Sources and Variables

The data of eTIMSS 2019 study conducted by the International Association for the Evaluation of Educational Achievement (IEA) was used in this study. This data are available to public use on IEA's website (Fishbein *et al*., 2021).

In the eighth grade level of eTIMSS 2019 mathematics item pool, there were a total of 208 stand-alone computerized items and 25 PSI items presented under three PSI tasks. There were a total of 16 booklets each of which was administered to a single student. The booklets 1-14 consisted of stand-alone eTIMSS items and booklets 15-16 contained PSI items. In each PSI booklet, there were a total of four tasks, two of them mathematics PSI tasks and other two were science tasks administered in two sessions where each session took around 45 minutes. The mathematics tasks were *Building, Robots, and Dinosaur Speed* of which the Building task was combined with *Robots* and presented/administered in a single session (Mullis *et al*., 2021). In the task of *Building,* one item was a multiple-choice item and the remaining eight of them were constructed-response items. Similarly, *Robot* included four constructed-response items and *Dinosaur* included one selected-response and sixteen constructed-response items.

In this study, two items ("Water Tank A" [MQ12B05A] and "Water Tank B" [MQ12B05B]) in *Building* task were studied, both were constructed-response items (Mullis *et al*., 2021, p. 110). Item response theory item parameters in *Building* tasks vary between 0.617 and 1.779 for discrimination, 0.467 and 2.084 for difficulty parameters. For "Water Tank A" and "Water Tank B" items, item parameters were 1.390 and 1.472 for discrimination, 0.771 and 0.816 for difficulty parameters respectively (Fishbein *et al*., 2021). Omit rates of items in *Building* also varied between 0.7% and 17.6%, the omitting rates for "Water Tank A" and "Water Tank B"

were 7.9% and 10.2% respectively. The omitted responses in *Building*, after excluding students who omitted all the items in the task, were recoded as "incorrect".

In this research, the study variables were student responses [incorrect(0)-correct(1)] to "Water Tank A" and "Water Tank B" items of *Building* task, a dichotomous variable showing whether the student used calculator or not during response generation for "Water Tank A" and "Water Tank B" items [not used (0)-used(1)] and first plausible values calculated for students' mathematics achievement across item pools scaled to a distribution with an international mean of 500 and a standard deviation of 100.

2.2. Sample

The TIMSS program employs a complex sampling method to increase the representation of the student population in each participated country. TIMSS uses stratified two-stage cluster random sample design in which a sample of schools drawn at first stage and one or more intact classes of students drawn from the sampled schools at second stage taking into account the stratification of schools depending on each participated countries' territorial-demographic characteristics (e.g., regions of the country, public-private schools, urban-rural areas). One apparent benefit of sampling the intact classes rather than individuals is easing the data collection process in terms of time and resources, and another benefit is that TIMSS pays particular attention to students' curricular and instructional experiences, and these are typically organized on a classroom basis (Martin, et al., 2020).

Students from each anticipated country and benchmarking participants were planned to include for this study first. However, there were students excluded from the analyses of this study. First, students who did not have access (not reached) to all the items in the given test were excluded. Similarly, students who did not answer all the items in the *Building* task were excluded. Students who were considered as noneffortful respondents were excluded from the analysis. Finally, some countries and benchmarking participants were excluded due to not having enough students in each cell of the levels of the variables (Table A in Appendix). Sample size of students who were administered *Building* task in eTIMSS 2019 cycle were presented in Table 1. How nonefforful respondents were decided are clarified in next paragraphs.

Table 1. *Number of eighth grade students included in analysis from each country.*

Country	Original Sample Size in TIMSS Dataset	Final Sample Size for Analysis
Chinese Taipei	665	644
Georgia	356	314
Hong Kong SAR	434	411
Hungary	588	572
Korea, Rep. of	503	479
Lithuania	453	445
Norway	446	402
Qatar	448	422
Russian Federation	423	408
Türkiye	523	513
United Arab Emirates	2792	2629
United States	1083	1043
Ontario, Canada	433	418
Quebec, Canada	368	356
Abu Dhabi, UAE	1060	973
Dubai, UAE	681	662
Total	11256	10691

Common approach for deciding noneffortful respondents is utilizing item response time information collected during testing. However, response times for each item were not available in eTIMSS2019 data, rather screen times were available. Unfortunate for the analysis, some screens contain several items. Therefore, response times for each screen were examined in this study.

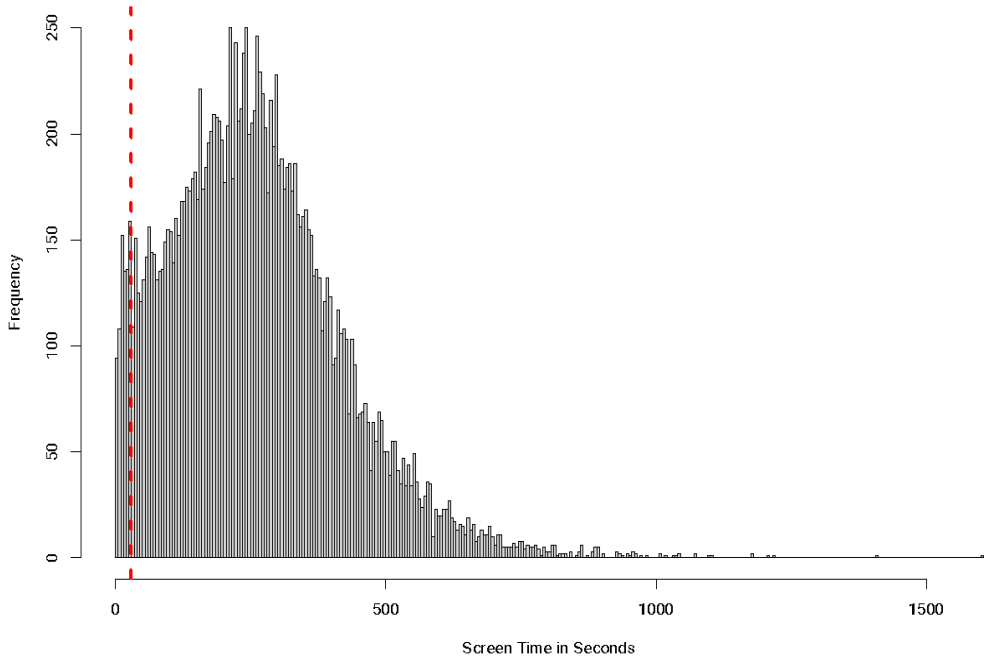
Table 2 shows screen time distributions for each screen consisting of items of the *Building* task. Screens completed less than a second implies noneffortful responding clearly. Previous researchers developed several methods to set a threshold of response time for filtering noneffortful respondents (e.g., Setzer et al., 2013; Ulitzsch et al., 2023; Wise, 2017; Wise & Gao, 2017). Among these methods, thresholds were set by using 3 or 5 seconds as common threshold across the items or calculating 10% of mean response time for an item with a maximum 10 seconds limitation (Wise et al., 2004; Wise & Ma, 2012). Though, the items in those studies were traditional item types (e.g., multiple-choice items) and response time distributions were available for each item.

Table 2. *Screen time distributions including items for "Building" task (in seconds).*

Screen	Min.	5th Quantile	25th Quantile	Mean	Median	75th Quantile	95th Quantile	Max.
Screen 2-Building Size	0.29	13.86	28.41	54.62	43.51	66.83	130.06	1194.93
Screen 3-Roof	0.21	13.20	37.70	71.59	59.12	91.44	165.94	771.99
Screen 4-Constructing the Walls	0.18	48.18	113.76	187.92	167.57	235.50	390.58	1550.70
Screen 5-Painting the Walls	0.29	31.28	98.39	170.10	152.46	220.39	365.54	1056.55
Screen 6-Water Tank	0.18	30.92	143.12	259.34	244.20	350.55	545.48	1600.26

The items in this study were part of a more complex problem solving task. There were three items on Screen 6, two of which were constructed-response items. Additionally, reading the instructions and the items in Screen 6 that require calculations can make the screen response time longer on average compared to other screens (Table 2). Considering 10 second-threshold in previous research contained relatively longer and complex items (Setzer et el., 2013), 30 seconds for a total of 3 items in Screen 6 were used as a threshold for screen response time. Screen time distribution given in Figure 1 also showed a "bump" on response time frequency occurred during the first 30 seconds that may be a sign of noneffortful responses of students (Schnipke, 1995). Therefore, students who spent less than 30 seconds on Screen 6 were excluded for eliminating noneffortful respondents.

Figure 1. *Screen time spent by the students showing the thresholds of 30 seconds.*



2.3. Data Analysis

The study has a cross-sectional design where strength of associations between dependent and independent variables were examined. Data analysis was conducted on R programming language environment (R Core Team, 2022, v.4.2.2) by modifying the relevant intsvy R package functions (Caro & Biecek, 2022, v.2.6).

2.3.1. *Sources of uncertainty and sampling variances*

The eTIMSS 2019 item pool contains 171 items with additional 29 PSI items in the fourth grade level item pool. Similarly, there were 206 items and 25 PSI items in the eighth grade level item pool. However, administering the entire item pool to each student would resulted in a burden of testing time. Instead, TIMSS uses matrix-sampling assessment design where each student is administered only a subset of items comparable through a common core of items. Based on the matrix-sampling approach, items were divided into 16 booklets where each item appeared in two booklets that allowed linking between booklets (Martin et al., 2017).

Matrix-sampling approach eases the testing process but it costs some variance and uncertainty in parameter estimates. One source of uncertainty is generalizing analysis results obtained from a student sample to the population of students called sampling variance, and second source of uncertainty is estimating students achievement scores from a sample of items called imputation variance (Foy & LaRoche, 2020).

2.3.1.1. Sampling Variance. The data were collected from national samples of students drawn once; therefore, how well the sample represents the target population is a crucial aspect of the analysis findings. As a result, sampling variance that also implies how well the sample represents the target population was computed and included during the analysis. The approach used for computing sampling variance in TIMSS 2019 was Jackknife Repeated Replication [(JRR), Foy & LaRoche, 2020].

2.3.1.2. Imputation Variance. In addition to sampling variance, as stated earlier, another variability is observed due to the fact that the student achievement is estimated by a subset of items instead of the entire item pool due to matrix-sampling assessment design. Students' achievement scores were generalized to the entire item pool by five plausible values (PV1-PV5)

computed by an imputation model. As a result, variation due to imputation procedures is observed in student achievement scores.

In summary, total variance in student achievement scores is obtained by summing JRR sampling variance and imputation variance; overall standard error for achievement estimations of each country is the square root of total variance computed for each country.

2.3.2. *Logistic Regression Models*

In order to answer the research questions, three binary logistic models were run in all of which the dependent variables *P* were probability of giving a correct response to the item. Models were given as follows:

$$\text{Model1: } log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_{PV1}$$

$$\text{Model2: } log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_{PV1} + \beta_{Calculator}$$

$$\text{Model3: } log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_{PV1} + \beta_{Calculator} + \beta_{PV1}{:}\beta_{Calculator}$$

In each of these models, $log\left(\frac{P}{1-P}\right)$ represents the natural logarithm of odds ratio (OR) of giving correct response, β_0 represents the intercept, β_{PV1} represents regression coefficient for first plausible value (PV1), and $\beta_{Calculator}$ represents the difference between reference group (non-calculator users) and focal group (calculator users) in the dependent variable.

In Model 1, PV1 was included as an independent variable only. In Model 2, a dichotomous variable that indicates the status of calculator use was added as another independent variable. In Model 3, two independent variables and their interaction effects were included. Since dependent variable was a dichotomous variable, estimated regression coefficients were associated to the change in log-odds of giving correct response with one-unit change in β_{PV1} and $\beta_{Calculator}$ and in their interaction term when controlling the effect of other independent variables.

For each logistic regression model, nested models were compared by chi-square difference tests (Tables 3-4). Additionally, McFadden R^2 as an approximation of the proportion of variance explained by independent variables (Smith & McKenna, 2013) and Akaike Information Criterion [(AIC); Bozdogan, 1987) were computed. These statistics were reported in Tables 3-4 and used for model comparison.

Ignoring the sampling and imputation variances while running logistic regression models can lead to bias in the estimation of standard errors and confidence intervals that may also cause incorrect interpretation of the results. Therefore in this study, total student weights (TOTWGT) and Jacknife replication values (JKZONE, JKREP) and first plausible values (PV1) were used to take into account the sampling variances and uncertainties.

3. RESULTS

Model comparison results for each fitted logistic regression models for “Water Tank A” and “Water Tank B” items were provided under M1-M3 columns where each represents model 1 through model 3 in Table 3 and Table 4. As seen in these tables, chi-square difference tests were examined and observed that for all the countries and benchmarking participants, model 2 had better model-data fit when compared to model 1. Similarly, adjusted McFadden R^2 values and AIC values showed that model 2 had a better fit with a higher proportion of explained variance and lower AIC values respectively when compared to model 1.

Next, model 2 and model 3 were compared. Accordingly, chi-square tests for “Water Tank A” showed that adding the interaction effect in model 3 provided a significant improvement for

Hong Kong SAR, Norway, Qatar, Türkiye, UAE, United States, Quebec-Canada and Abu Dhabi-UAE when compared to model 2 (α= 0.05). Similarly, chi-square tests for "Water Tank B" showed that adding the interaction effect in model 3 provided a significant improvement for Georgia, Hong Kong SAR, Republic of Korea, Norway, Qatar, Türkiye, UAE, United States, Ontario-Canada and Abu Dhabi-UAE when compared to model 2 (α= 0.05). AIC values were also lower for these countries specified above for both items, though adjusted McFadden R^2 values did not seem provide a larger proportion of variance explained in model 3 when compared to model 2. These findings suggest that using digital calculators are positively associated with higher odds of giving correct responses for both items above and beyond students` mathematics achievement scores conditional on students' mathematics achievement scores; however, odds-ratio coefficients vary across the status of calculator use for some of the countries.

Table 3. *Logistic regression model comparison statistics for "Water Tank A".*

Country	Chi-Square Test		Adjusted McFadden R^2			AIC		
	M1-M2	M2-M3	M1	M2	M3	M1	M2	M3
Chinese Taipei	< 0.001	0.058	0.31	0.37	0.37	630.71	573.48	572.54
Georgia	< 0.001	0.323	0.37	0.43	0.43	165.03	147.70	148.19
Hong Kong SAR	< 0.001	0.013	0.31	0.32	0.33	399.50	394.37	389.54
Hungary	< 0.001	0.896	0.34	0.45	0.44	606.10	508.64	510.57
Korea, Rep. of	< 0.001	0.793	0.37	0.45	0.45	404.80	352.94	354.83
Lithuania	< 0.001	0.643	0.33	0.36	0.36	392.67	374.38	376.28
Norway	< 0.001	0.027	0.23	0.30	0.31	445.82	408.63	402.50
Qatar	< 0.001	0.011	0.37	0.44	0.45	328.75	293.85	288.10
Russian Fed.	< 0.001	0.120	0.33	0.35	0.34	404.14	393.85	394.18
Türkiye	< 0.001	0.017	0.41	0.50	0.51	263.21	220.95	217.25
UAE	< 0.001	0.023	0.31	0.33	0.33	2335.08	2270.70	2263.37
United States	< 0.001	0.014	0.31	0.32	0.32	1097.31	1076.01	1074.34
Ontario, Canada	< 0.001	0.967	0.25	0.27	0.27	436.04	425.43	427.45
Quebec, Canada	< 0.001	0.047	0.19	0.21	0.22	390.93	378.93	376.92
Abu Dhabi, UAE	< 0.001	0.009	0.32	0.33	0.34	727.07	717.69	711.10
Dubai, UAE	< 0.001	0.716	0.28	0.31	0.30	671.04	651.40	653.47

Note1. UAE: United Arab Emirates.
Note2. ChiSquare difference test was evaluated at α=0.05 level.
Note3. Model 2 was adopted for Chinese Taipei, Georgia, Hungary, Republic of Korea, Lithuania, Russian Federation, Ontario-Canada, Dubai-UAE.
Note4. Model 3 was adopted for Hong Kong SAR, Norway, Qatar, Türkiye, UAE, United States, Quebec-Canada and Abu Dhabi-UAE.

To provide a further demonstration, how calculator use was associated with higher probability of giving correct responses conditional on students' mathematics achievement scores were presented with plots. As shown in Figure 2 and Figure 3, the group of students who used calculators in both items had higher probability of giving correct responses when compared to the group of students who did not use calculators having the same mathematics scores on average. The statistically significant interaction effects between calculator use and mathematics scores for the countries who were listed above can be observed in Figure 2 and Figure 3.

Table 4. *Logistic regression model comparison statistics for "Water Tank B".*

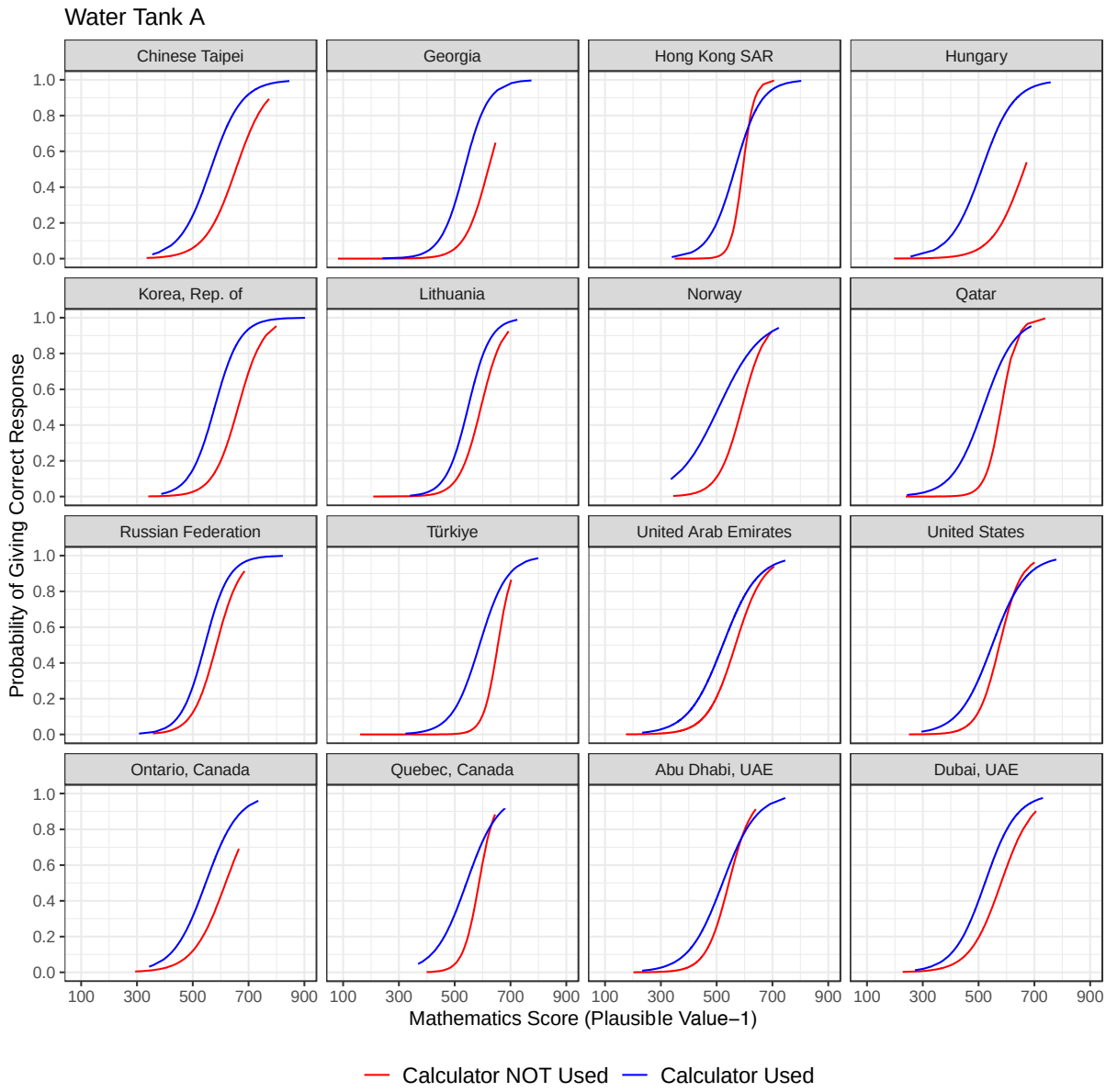
Country	Chi-Square Test		Adjusted McFadden R^2			AIC		
	M1-M2	M2-M3	M1	M2	M3	M1	M2	M3
Chinese Taipei	< 0.001	0.957	0.31	0.37	0.37	619.82	600.83	602.83
Georgia	0.020	0.003	0.37	0.43	0.43	154.22	151.35	141.30
Hong Kong SAR	0.003	0.041	0.31	0.32	0.33	417.55	410.71	407.57
Hungary	< 0.001	0.988	0.34	0.45	0.44	575.59	540.64	542.64
Korea, Rep. of	< 0.001	0.029	0.37	0.45	0.45	436.55	408.39	406.62
Lithuania	< 0.001	0.946	0.33	0.36	0.36	402.04	381.53	383.52
Norway	< 0.001	< 0.001	0.23	0.30	0.31	448.06	397.20	382.77
Qatar	0.001	0.021	0.37	0.44	0.45	310.68	299.97	293.40
Russian Fed.	0.002	0.693	0.33	0.35	0.34	365.27	355.74	357.65
Türkiye	0.004	0.012	0.41	0.50	0.51	240.30	234.91	230.84
UAE	< 0.001	< 0.001	0.31	0.33	0.33	2110.97	2060.75	2047.92
United States	< 0.001	< 0.001	0.31	0.32	0.32	1022.31	994.48	980.68
Ontario, Canada	0.001	0.033	0.25	0.27	0.27	472.51	461.64	457.90
Quebec, Canada	0.001	0.097	0.19	0.21	0.22	373.40	363.48	362.18
Abu Dhabi, UAE	0.001	< 0.001	0.32	0.33	0.34	693.35	683.52	673.70
Dubai, UAE	< 0.001	0.559	0.28	0.31	0.30	633.39	612.09	613.63

Note1. UAE: United Arab Emirates.
Note2. ChiSquare difference test was evaluated at α=0.05 level.
Note3. Model 2 was adopted for Chinese Taipei, Hungary, Lithuania, Russian Federation, Quebec-Canada, Dubai-UAE.
Note4. Model 3 was adopted for Georgia, Hong Kong SAR, Republic of Korea, Norway, Qatar, Türkiye, UAE, United States, Ontario-Canada and Abu Dhabi-UAE.

As seen in the Figures, for the participating countries and benchmarking that show a significant interaction effect, regression coefficients between student' mathematics scores and odds of giving correct response were not equal across the calculator users or non-users. Therefore, students who did not use the on-screen calculator and who had a score of 600 or higher had similar or even higher probabilities of giving the correct responses when compared to the ones who did not use the calculator. The authors note that the statistical coefficients are also a function of sample size and observed significant interaction effect may be due to relatively larger sample size in countries such as United Arab Emirates, United States or Abu Dhabi. Additionally, the prediction of probabilities for giving correct responses are limited to the range of the predictor data on x-axes.

How the findings of this study are consistent with the findings of the previous research were discussed in the next section. The impact of current study findings to the educational measurement literature and implications for the computerized item development were presented in the Discussion section.

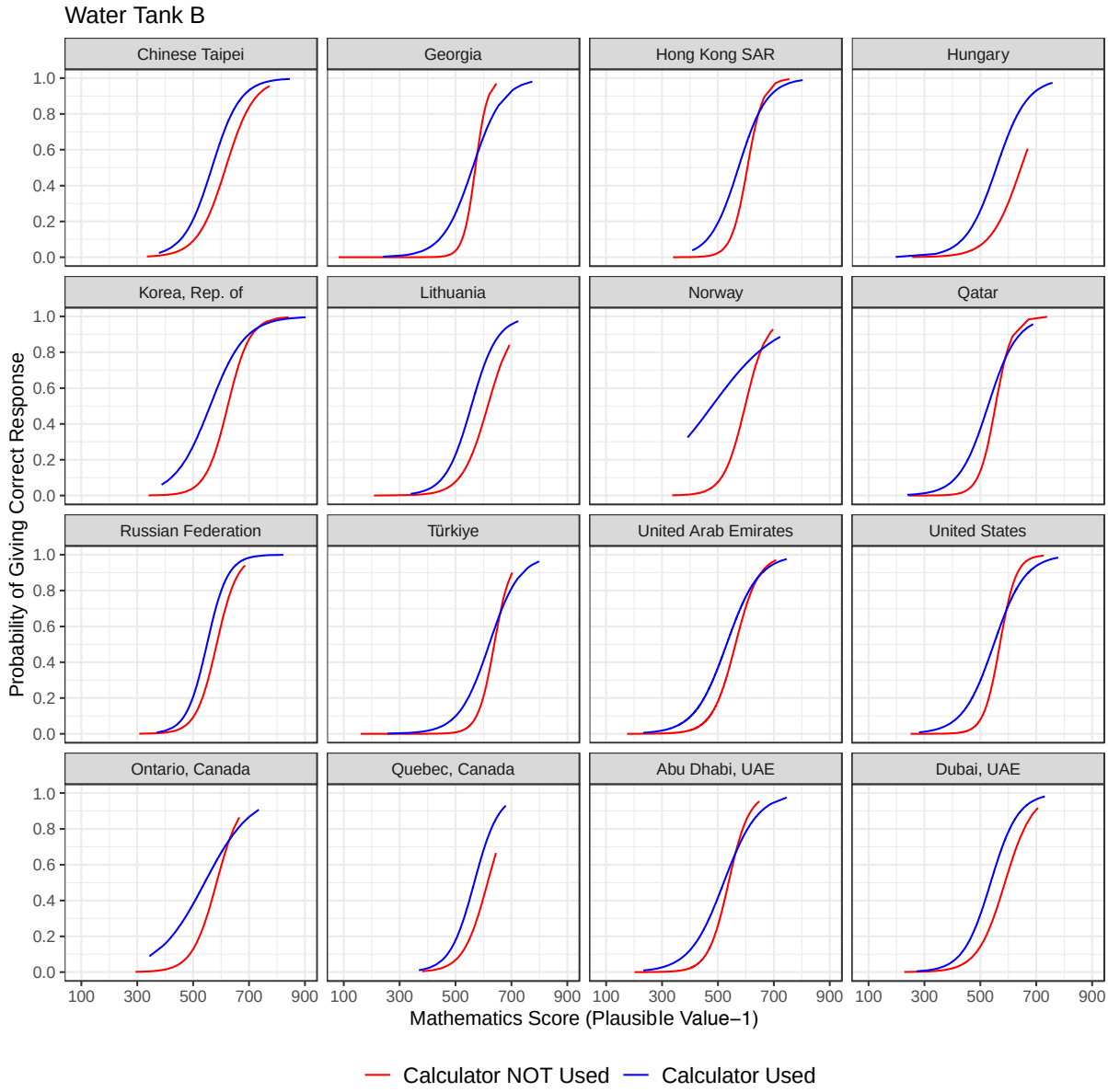
Figure 2. *Predicted probabilities of giving correct response to "Water Tank A" conditional on mathematics achievement and calculator use based on adopted logistic regression model.*



Note1. Model 2 was adopted for Chinese Taipei, Georgia, Hungary, Republic of Korea, Lithuania, Russian Federation, Ontario-Canada, Dubai-UAE.

Note2. Model 3 was adopted for Hong Kong SAR, Norway, Qatar, Türkiye, UAE, United States, Quebec-Canada and Abu Dhabi-UAE.

Figure 3. *Predicted probabilities of giving correct response to "Water Tank B" conditional on mathematics achievement and calculator use based on adopted logistic regression model.*



Note1. Model 2 was adopted for Chinese Taipei, Hungary, Lithuania, Russian Federation, Quebec-Canada, Dubai-UAE.
Note2. Model 3 was adopted for Georgia, Hong Kong SAR, Republic of Korea, Norway, Qatar, Türkiye, UAE, United States, Ontario-Canada and Abu Dhabi-UAE.

4. DISCUSSION and CONCLUSION

Discussions on the use of calculators have become a research topic in recent years at the point of designing it as a tool that can be used during learning and assessments, even as a digital tool that students can use on screen for computerized tests. Early research findings showed that calculator use can improve computational skills of students with average ability and have no adverse effects on the computational skills of the low and the high ability students (Brolin & Bjork, 1992: Hembree & Dessart, 1986; Hembree & Dessart, 1992.

Additionally, studies reveal that the use of calculators supports students during assessments. To solve a problem, the students must understand the problem, decide which problem-solving strategy is appropriate, carry out the strategy, and determine the solution. Therefore, calculators can contribute complex computing processes while students can spend more time on thinking

and developing a strategy (NCTM, 2015; Sparrow et al., 1994; Vasquez & McCabe, 2002). Previous studies in which large-scale assessment data were used showed that students who used calculators for mathematics problem solving items had significantly higher test scores than the students who did not use them (Mullis et al., 2021: Walcott & Stickles, 2012).

Current study results are parallel to the literature that promotes the use of calculators as a supportive tool during assessments. Current study findings showed that students who used the on-screen calculator had significantly higher probability of giving correct response above and beyond their mathematics achievements. As a result, it is suggested that the test and item developers should consider adding the on-screen calculator tool to the item as part of the innovations in TE items if test specifications and construct being measured allow. With the lights of the current study findings, more structured research is needed to collect further validity evidence regarding on-screen calculators.

The research findings also suggest that for some of the participating countries and benchmarkings, the interaction effect between student' mathematics scores and calculator use status was significant. This means that the odds of giving correct response were not equal across the calculator users or non-users in some of the countries. This observation may be related to the countries' education programs and students' familiarity and being used to the calculators in solving the mathematics problems. For instance, previous research indicated that the majority of the eighth grade students in participating European countries were allowed to use calculators approximately half or more than half of lessons to solve complex problems, do routine computations, and check answers (Eurydice, 2011). Considering the European students' potential familiarity with calculators, the proportion of students who answered the items correctly among the students who come from the European countries and did not use the calculators is extremely small is not surprising (Table A in Appendix). Similarly, even though Singapore is a high achieving country, the proportion of students who answered the items correctly is small among the students who did not use the calculator that may be related to the students' familiarity with using calculators starting from fifth grade (Koay, 2006; Mullis et al., 2016). Though, why a significant interaction effect was found in only some of the countries require further review and research.

This study is not without limitations. In this study, two items given under PSI tasks of TIMSS study were studied and the role of calculator use in other items in eTIMSS 2019 could not be studied due to the fact that such process data were available only for those two items in publicly available data. Yet, the findings of this study serve as preliminary findings and the content and context of the study can be expanded with more detailed process data regarding the use of calculators or other digital tools (e.g., ruler) with TIMSS data or any other TE items data.

There were 27 countries and benchmarking participants in eTIMSS 2019; however, data analysis was completed with students from only 16 countries and benchmarking participants in this study. The reason for this situation was that there were not enough students in each cell of the study variables (Table A in Appendix). Future research can examine if there are specific characteristics of these excluded countries that are relevant to using calculators during mathematics classrooms and assessments. Additionally, as prediction plots in Figure 2 and Figure 3 indicate, calculator use does not impact the probability of giving correct response at a fixed rate for some countries, rather high ability students may not need to use them as their problem solving processes. Therefore, future research can also examine what characteristics of their education system are associated with such findings for these countries that may inform the item and test development processes for country-specific assessments or cross-cultural assessments due to fairness.

Acknowledgments

The authors would like to thank the blind reviewers for their valuable feedback and suggestions. We also would like to thank to Emine Özdemir for the help in editing and proofreading the manuscript.

Declaration of Conflicting Interests and Ethics

The authors declare 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 authors.

Contribution of Authors

Ozge Ersan: Literature review, research design, data processing, data analysis, interpretations, writing and revising the manuscript. **Burcu Parlak:** Literature review, research design, interpretations, writing and revising the manuscript.

Orcid

Ozge Ersan https://orcid.org/0000-0003-0196-5472
Burcu Parlak https://orcid.org/ 0000-0001-7515-7262

REFERENCES

Betts, J., Muntean, W., Kim, D., & Kao, S. (2022). Evaluating different scoring methods for multiple response items providing partial credit. *Educational and Psychological Measurement, 82*(1), 151–176. https://doi.org/10.1177/0013164421994636

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370. https://doi.org/10.1007/BF02294361

Brolin, H., & Bjork, L-E (1992). Introducing calculators in Swedish schools. In J.T. Fey & C. R. Hirsch (Eds.), *Calculators in mathematics educationi* (pp. 226–232). Reston, VA: National Council of Teachers of Mathematics.

Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research & Evaluation, 22*(1).

Caro D.H. & Biecek P. (2022). *intsvy: International Assessment Data Manager*. R package version 2.6. https://CRAN.R-project.org/package=intsvy

Clyne, C.M. (2015). *The effects of different scoring methodologies on item and test characteristics of technology-enhanced items* [unpublished doctoral dissertation]. University of Kansas, Lawrence, Kansas. https://kuscholarworks.ku.edu/bitstream/handle/1808/21675/Clyne_ku_0099D_14314_DATA_1.pdf?sequence=1

Cohen, A.S. & Kim, S. (1992). Detecting calculator effects on item performance. *Applied Measurement in Education, 5*(4), 303-320. https://doi.org/10.1207/s15324818ame0504_2

Dick, T. (1988). The continuing calculator controversy. *Arithmetic Teacher*, 37–41.

Ellington, A.J. (2003). A meta-analysis of the effects of calculators on students' achievement and attitude levels in precollege mathematics classes. *Journal for Research in Mathematics Education, 34*(5), 433–463. https://doi.org/10.2307/30034795

Eurydice (2011). *Mathematics education in Europe: common challenges and national policies.* http://keyconet.eun.org/c/document_library/get_file?uuid=e456b461-d3cd-4bd5-aabc-2cae2d4bfaf9&groupId=11028

Fishbein, B., Foy, P., & Yin, L. (2021). *TIMSS 2019 User Guide for the International Database (2nd ed.).* TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation

of Educational Achievement (IEA). https://timssandpirls.bc.edu/timss2019/international-database

Foy, P., & LaRoche, S. (2020). Estimating standard errors in the TIMSS 2019 results. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 14.1–14.60). TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Hembree, R., & Dessart, D.J. (1986). Effects of hand-held calculators in precollege mathematics education: A meta-analysis. *Journal for Research in Mathematics Education*, *17*(2), 83–99. https://doi.org/10.2307/749255

Hembree, R., & Dessart, D.J. (1992). Research on calculators in mathematics education. In J.T. Fey & C.R. Hirsch (Eds.), *Calculators in mathematics education: 1992 NCTM Yearbook* (pp. 23–32). Reston, VA: The National Council of Teachers of Mathematics.

Hopkins, M.H. (1992). The use of calculators in assessment of mathematics. In T. Fey & C. R. Hirsch (Eds.), *Calculators in mathematics education: 1992 NCTMYearbook* (pp. 158–166). Reston, VA: The National Council of Teachers of Mathematics.

Impara, J.C., & Foster, D. (2006). Question and test development strategies to minimize test fraud. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 91–114). Lawrence Erlbaum Associates.

Koay, P.L. (2006). Calculator use in primary school mathematics: A Singapore perspective. *The Mathematics Educator*, *9*(2), 97-111.

Lorié, W. (2016). Automated scoring of multicomponent tasks. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (p. 627–658). IGI Global. https://doi.org/10.4018/978-1-4666-9441-5.ch024

Loyd, B. H. (1991). Mathematics test performance: The effects of item type and calculator use. *Applied Measurement in Education*, *4*(1), 11–22.

Ludwig, S., & Rausch, A. (2022). The relationship between problem-solving behaviour and performance – Analysing tool use and information retrieval in a computer-based office simulation. *Journal of Computer Assisted Learning*, 1-27. https://doi.org/10.1111/jcal.12770

Martin, M.O., Mullis, I.V.S., & Foy, P. (2017). TIMSS 2019 Assessment Design. In I.V.S. Mullis, & M.O. Martin (Eds.), *TIMSS 2019 Assessment Frameworks* (pp. 79–92). TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Martin, M.O., von Davier, M., & Mullis, I.V.S. (2020). *Methods and Procedures: TIMSS 2019 Technical Report.* TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Mullis, I.V.S., Martin, M.O. (2017). *TIMSS 2019 Assessment Frameworks*. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA). https://timss2019.org/psi/

Mullis, I.V.S., Martin, M.O., Fishbein, B., Foy, P., & Moncaleano, S. (2021). *Findings from the TIMSS 2019 problem solving and inquiry tasks*. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA). https://timss2019.org/psi/

Mullis, I.V.S., Martin, M.O., Goh, S., & Cotter, K. (Eds.) (2016). *TIMSS 2015 Encyclopedia: Education Policy and Curriculum in Mathematics and Science.* Retrieved from Boston College, TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA). http://timssandpirls.bc.edu/timss2015/encyclopedia/

National Council of Teachers of Mathematics (2015). *Calculation Use in Elementary Grades.* https://www.nctm.org/Standards-and-Positions/Position-Statements/Calculator-Use-in-Elementary-Grades

Parshall, C.G., Harmes, J.C., Davey, T., & Pashley, P.J. (2010*).* Innovative items for computerized testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Elements of adaptive testing* (pp. 215–230). Springer. https://doi.org/10.1007/978-0-387-85461-8

Plunkett, S. (1978). Decomposition and all that rot. *Mathematics in Schools, 8*(3), 2–5.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Russell, M., & Moncaleano, S. (2019) Examining the use and construct fidelity of technology-enhanced items employed by K-12 testing programs. *Educational Assessment, 24*(4), 286–304. https://doi.org/10.1080/10627197.2019.1670055

Salles, F., Dos Santos, R., & Keskpaik, S. (2020). When didactics meet data science: process data analysis in large-scale mathematics assessment in France. *Large Scale Assessment in Education 8*(7). https://doi.org/10.1186/s40536-020-00085-y

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: a framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment*, *4*(6).

Schnipke, D.L. (1995). *Assessing speededness in computer-based tests using item response times* [Unpublished doctoral dissertation]. Johns Hopkins University, Baltimore, MD.

Setzer, J.C., Wise, S.L., van den Heuvel, J.R., & Ling, G. (2013) An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education, 26*(1), 34–49. https://doi.org/10.1080/08957347.2013.739453

Sireci, S.G. & Zenisky, A.L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (p. 329–348). Lawrence Erlbaum Associates.

Smith, T.J., & McKenna, C.M. (2013). A comparison of logistic regression pseudo R^2 indices. *General Linear Model Journal, 39*(2), 17–26. http://www.glmj.org/archives/GLMJ_2014v39n2.html

Sparrow, L., Kershaw, L., & Jones, K. (1994). *Issues in primary mathematics education: calculators: research and curriculum implications.* Perth, Australia: Mathematics, Science & Technology Education Centre, Edith Cowan University.

Tarr, J.E., Uekawa, K., Mittag, K.C., & Lennex, L. (2000). A comparison of calculator use in eighth-grade mathematics classrooms in the United States, Japan, and Portugal: Results from the Third International Mathematics and Science Study. *School Science and Mathematics*, *100*(3), 139–150. https://doi.org/10.1111/j.1949-8594.2000.tb17249.x

Ulitzsch, E., Domingue, B.W., Kapoor, R., Kanopka, K. and Rios, J.A. (2023). A probabilistic filtering approach to non-effortful responding. *Educational Measurement: Issues and Practice*. Advanced online publication. https://doi.org/10.1111/emip.12567

Vasquez, S., & McCabe, T.W. (2002). The effect of calculator usage in the learning of basic skills. *Research and Teaching in Developmental Education*, *19*(1), 33–40.

Walcott, C., Stickles, P.R. (2012). Calculator Use on NAEP: A look at fourth- and eighth-grade mathematics achievement. *School Science and Mathematics*, *112*(4), 241–254. https://doi.org/10.1111/j.1949-8594.2012.00140.x

WIDA (n.d.). (2023) *2022-2023 Accessibility & accommodations Manual.* https://wida.wisc.edu/sites/default/files/resource/Accessibility-Accommodations-Manual.pdf

Williams, D. (1987). Using calculators in assessing mathematics achievement. *Arithmetic Teacher, 34*(2), 21-23.

Wise, S.L. (2017), Rapid-guessing behavior: its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*, 52-61. https://doi.org/10.1111/emip.12165

Wise, S.L., & Gao, L. (2017) A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education, 30*(4), 343-354, https://doi.org/10.1080/08957347.2017.1353992

Wise, S.L., Kingsbury, G.G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Wise, S.L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wolfe, E.W. (2010). What impact does calculator use have on test results? *Test, Measurement & Research Services Bulletin, 14*, 1–6.

APPENDIX

***Table A**. Number of Students Depending on Their Use of Calculator and Giving Correct Responses.*

Country	Water Tank A			Water Tank B		
	Calc. Users	n	Number of Students Correctly Responded	Calc. Users	n	Props. of Students Correctly Responded
Chile	0	205	1	0	229	4
	1	173	60	1	149	51
Chinese Taipei	0	275	60	0	319	106
	1	369	253	1	325	229
England	0	135	1	0	178	6
	1	245	122	1	202	107
Finland	0	211	0	0	256	4
	1	311	102	1	266	107
France	0	130	0	0	175	3
	1	244	85	1	199	93
Georgia	0	203	11	0	220	15
	1	111	42	1	94	33
Hong Kong SAR	0	72	13	0	105	20
	1	339	217	1	306	193
Hungary	0	198	14	0	230	20
	1	374	257	1	342	205
Israel	0	160	2	0	206	9
	1	239	99	1	193	103
Italy	0	135	7	0	162	10
	1	255	112	1	228	101
Korea, Rep. of	0	216	36	0	235	60
	1	263	182	1	244	171
Lithuania	0	185	37	0	204	31
	1	260	128	1	241	120
Malaysia	0	197	6	0	201	13
	1	695	257	1	691	258
Norway	0	234	50	0	250	50
	1	168	101	1	152	95
Portugal	0	120	3	0	145	4
	1	256	102	1	231	89
Qatar	0	208	11	0	228	21
	1	214	85	1	194	74
Russian Federation	0	118	26	0	141	28
	1	290	172	1	267	164
Singapore	0	43	8	0	58	13
	1	579	462	1	564	469
Sweden	0	103	2	0	137	3
	1	247	118	1	213	95
Türkiye	0	364	13	0	385	22

	1	149	52	1	128	38
United Arab Emirates	0	1124	166	0	1335	198
	1	1505	664	1	1294	548
United States	0	254	31	0	314	38
	1	789	396	1	729	373
Ontario, Canada	0	97	18	0	111	22
	1	321	168	1	307	175
Quebec, Canada	0	65	13	0	89	14
	1	291	171	1	267	138
Moscow, Russian Fed.	0	71	6	0	94	9
	1	357	247	1	334	236
Abu Dhabi, UAE	0	488	65	0	564	80
	1	485	182	1	409	161
Dubai, UAE	0	197	43	0	244	54
	1	465	280	1	418	237

Note. 0: did not use calculator, 1: used calculator.