Boston College
Lynch School of Education and Human Development

Department of
Measurement, Evaluation, Statistics, and Assessment

EXAMINING THE COMPARATIVE MEASUREMENT VALUE OF TECHNOLOGY-ENHANCED ITEMS

Dissertation
by

SEBASTIAN MONCALEANO

submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

MAY 2021

Examining the Comparative Measurement Value of Technology-Enhanced Items

Sebastian Moncaleano

Dissertation Chair: Dr. Michael Russell

Abstract

The growth of computer-based testing over the last two decades has motivated the creation of innovative item formats. It is often argued that technology-enhanced items (TEIs) provide better measurement of test-takers' knowledge, skills, and abilities by increasing the authenticity of tasks presented to test-takers (Sireci & Zenisky, 2006). Despite the popularity of TEIs in operational assessments, there remains little psychometric research on these innovative item formats. Claims regarding their potential to provide better measurement are seldomly explored. This dissertation adds to this limited body of research by developing theory and proposing a methodology to compare TEIs to traditional item formats.

This study investigated how to judge the comparative measurement value (CMV) of two drag-and-drop technology-enhanced formats (classification and rank-ordering) relative to stem-equivalent multiple-choice items. Items were administered to a sample of adults and results were calibrated using a 2-parameter logistic IRT model. Moreover, the utility of the TEIs was evaluated according to the TEI Utility Framework (Russell, 2016).

Four indicators were identified as the most valuable characteristics to judge CMV and then combined into a hierarchical decision protocol. When applied, this protocol provides a CMV judgment and a recommendation of the preferred item format. Applying the protocol to the items revealed that most TEIs examined in this study showed

decreased CMV, indicating that in a real-life scenario the multiple-choice format would be favored for most of these item pairs. Recommendations for the use of the CMV protocol and directions of future related research are discussed.

Acknowledgments

This dissertation is the result of several years of work and it would not have been possible without the support of many.

First, thanks to my committee. To my dissertation chair, Dr. Michael Russell, whose mentorship over the past six years made my doctoral journey enriching and rewarding. To my readers Dr. Zhushan Li and Dr. Lillie Albert for providing feedback and guidance when I needed it most. And to all three of them for their support and understanding when my project was forced to evolve due to the ongoing COVID-19 pandemic.

To my best friend and colleague, Katherine Reynolds, for being a constant companion through this ride; who in the ups and the downs was always there for me, pushing me to be the best I could. Also, as the only person outside of my committee to read every word of this dissertation, for having the love and patience to do so and provide her thoughtful feedback.

Thanks also to my family–Mom (Gloria), Dad (Carlos), and my sister (Pau). Moving away to pursue this degree was hard and despite the distance I always felt supported. To my support network without whom I would not have gotten this far. Thank you to Angela Duarte, Juan Pablo Orozco, Carolina Calkins, and Kristopher Cannon who always believed I'd get here and were always ready to share some words of encouragement. My gratitude will always be with my college mentors who fostered my love and interest in education and academic research, Maria Figueroa, Oscar Bernal, and Maria Teresa Gomez. Also, thank you to all my friends who invested their time to help me refine my data collection instrument.

Finally, I dedicate this dissertation to my husband, Luke, to whom I owe his unwavering love and support. From 1:00 AM brainstorming sessions to keeping me sane and entertained through this pandemic, his constant care was paramount to make this dissertation real.

Foreword

This dissertation was motivated by a strong belief that technology-enhanced items should face more scrutiny regarding the contexts in which they should be used. All too often we, as test developers and item writers, ask *How can this item be more interactive?* rather than asking *Would a technology-enhanced interaction allow assessing more accurately the intended construct?* Grounded in the Evidence Centered Design framework (Mislevy, 2003), accessing the targeted constructs should be a priority and test developers should not be distracted by the attractive opportunities technological innovations provide. This dissertation presents the Comparative Measurement Value Protocol which is intended as a tool to inform decisions regarding when technology-enhanced items provide better measurement compared to stem-equivalent multiple-choice items. This protocol provides evidence-based rationales that will inform validity arguments of operational assessment programs on the usage of common technology-enhanced formats.

The study described in this dissertation was conducted in the summer of 2020, at the height of the COVID-19 pandemic. This study was planned and ready to be executed in the spring of that same year in 8th-grade classrooms. However, as schools closed to prevent the spread of the disease this study had to be re-designed. As described throughout this work, relying on Amazon's MTurk to gather data posed multiple challenges. Despite these hurdles, this was the first dissertation in the Measurement, Evaluation, Statistics, and Assessment department at Boston College to overcome the limitations to data collection posed by the pandemic. Let this work be a testament of the impact of this world-wide event on educational research.

Table of Contents

Tables

Figures

Chapter 1 - Introduction

The last two decades have seen a significant increase in the use of computer-based tests (CBTs) as part of large-scale assessment programs. The growing prevalence of computer-based testing has motivated the creation of innovative item formats to improve the way tests assess examinees. It is often argued that technology-enhanced items (TEIs) provide a better measurement of test-takers' knowledge, skills, and abilities by increasing the authenticity of tasks performed while responding to a test item (Sireci & Zenisky, 2006). The goal of increasing authenticity has led to the development of several TEI formats and response interaction spaces and the subsequent adoption of TEIs by several large-scale assessment programs. However, despite the popularity of TEIs in operational assessments, there remains little psychometric research on these innovative item formats; claims regarding their potential to provide better measurement of student achievement have been seldomly explored. Consequently, it is necessary to gather empirical data to evaluate these claims.

This dissertation aims to add to the limited body of research that has examined empirically the psychometric properties of TEIs and proposes a methodology to compare TEIs to traditional item formats. The present chapter describes the background and motivation behind this dissertation. This chapter begins with a review of the growth of computer-based tests in the United States followed by a brief introduction to technology-enhanced items. Next, the research objectives of this dissertation and an overview of the methodology employed are presented. The chapter concludes with a discussion of the significance of this study and a description of the content of future chapters.

The Growth of Computer-Based Testing in the United States

The use of computers to deliver educational tests was pioneered by Educational Testing Service (ETS) during the 1990s (Briel & Michel, 2014; Moncaleano & Russell, 2018). The launch of the world wide web enabled efficient digital communication and transfer of data. The testing industry capitalized on this efficiency to administer tests online at testing centers and to distribute score reports to academic institutions (Clarke et al., 2000). These benefits prompted ETS to transition the Graduate Record Exam (GRE) General Test from a paper-based format to a computer-based format in 1992. In subsequent years, ETS migrated several other tests to a digital format, including the Graduate Management Admission Test (GMAT) and the SAT I (Bennett, 1998). By the end of the decade, computer-based administration became the norm rather than the exception for tests provided by ETS.

In the early 2000s, simultaneous changes at the federal and state levels in the United States set the stage for the rise of computer-based testing in K-12 settings. At the federal level, growing concern about the lack of competitiveness of the American education system compared to other developed nations in the late 20th century led to the introduction of the No Child Left Behind Act (NCLB) of 2001. President George W. Bush highlighted the lack of growth in the National Assessment of Educational Progress (NAEP) and how the United States performed below its industrialized competitors in the Third International Mathematics and Science Study (TIMSS) to argue that increased accountability, and consequently increased testing, were key to improving the U.S. education system (Madaus et al., 2009). Among other elements, the NCLB law mandated states to test students in mathematics and reading in grades 3 through 8 and once in high

school (Klein, 2015). Concurrently, states began developing computer-based tests for large-scale administrations. In 2002, Oregon and South Dakota became the first states to implement computer-based tests while 10 additional states were piloting CBTs for future administration (Borja, 2002). The nexus between the sudden increase in federal testing requirements, the growing availability of computers, and the surge of interest in applications of technology to educational settings led to the rapid proliferation of state-wide computer-based tests throughout the United States in the following years. By 2003, 12 states had implemented digital tests (Edwards et al., 2003) and by 2006 the total had reached 22 states (Swanson, 2006). By the end of the decade, approximately half of the states (26) were using computers to deliver at least a portion of their annual state test (Blazer, 2010; Thurlow et al., 2010).

In September 2010, the U.S. Department of Education allocated $350 million to develop next-generation assessments through the Race to the Top Assessment (RTTA) program (Race to the Top Fund Assessment Program, 2010). The program provided funding for states to develop tests that "support and inform instruction, provide accurate information about what students know and can do, and measure student achievement against standards designed to ensure that all students gain the knowledge and skills needed to succeed in college and the workplace" (U.S. Department of Education, 2010, p. 2). The RTTA program led to the formation of six federally funded assessment consortia. The Partnership for Assessment of Readiness for College and Careers (PARCC) and the SMARTER Balanced Assessment Consortium (SBAC) were awarded grants by the RTTA program to develop digitally-delivered comprehensive Mathematics and English Language Arts assessments for all students except those with significant disabilities. The RTTA

program also funded two consortia to "develop next generation assessments for students with the most significant cognitive disabilities" (Educational Testing Service [ETS], 2016, p. 17): the Dynamic Learning Maps Alternate Assessment Consortium (DLM) and the Multistate Alternate Assessment Consortium (MSAA; formerly known as the National Center and State Collaborative, NCSC). In subsequent years, the RTTA program also awarded grants to two consortia to develop English language proficiency assessments. The WIDA collaborative was funded to develop the ACCESS for ELLs 2.0 assessment system in 2011 while the English Language Proficiency Assessment for the 21st Century (ELPA21) consortium was funded in 2012 (ETS, 2016). The funding provided by the RTTA program spurred wide-spread development of computer-based tests and introduced digital features that improved test accessibility for students with cognitive disabilities.

Several national and international assessment programs also transitioned or launched efforts to transition to digital platforms during the past two decades. NAEP began researching the possibility of transitioning its tests to computers in 2001 (Bennet et al., 2008) and eventually conducted a large-scale pilot for computer- and tablet-based tests in 2016 (National Assessment of Educational Progress, 2018). NAEP officially transitioned its mathematics and reading tests to digitally-based assessments in 2017 (Jewsbury et al., 2020). Internationally, the Programme for International Student Assessment (PISA) began transitioning to a computer-based platform in 2006 (Organization for Economic Co-operation and Development, 2010) and half the countries participating in TIMSS 2019 administered the assessment on computers or tablets (Fishbein et al., 2018; Martin et al., 2017; TIMSS & PIRLS International Study Center, 2019). Although some testing

programs remain dependent on paper-based administration, many local, state, national, and international assessment programs are now administered in a digital format.

Technology-Enhanced Items

Two of the most frequently cited benefits of computer-based educational assessments are: (a) efficient administration and scoring and (b) the use of innovative item formats to improve construct representation and fidelity (Sireci & Zenisky, 2006). Large-scale standardized paper-based tests often rely on selected-response (SR) items because their constrained nature allows testing a broad range of topics in a short amount of time as well as fast and efficient scoring. In turn, the administration of several SR items in a short amount of time generally improves two important measurement properties of the test, namely reliability and content representation (Haladyna & Rodriguez, 2013). Selected response items (SRIs), however, are criticized for being too constrained to assess some constructs at an appropriate level of depth and complexity (Madaus & O'Dwyer, 1999; Scully, 2017). Although some tests include constructed-response (CR) items, use of these items is generally limited because they require more time to answer compared to SRIs and increase scoring costs (Wendler & Walker, 2006).

Computer platforms allow standardized tests to include new item formats that capitalize on innovative response interactions to capture with more authenticity test-takers' understanding of assessed constructs (Russell, 2016). Items in digital environments may be enhanced in three main ways: (a) by enhancing the prompt, (b) by enhancing the way test-takers interact with the item to produce a response, or (c) by enhancing both the prompt and the response interaction space. For example, the Test of English as a Foreign Language (TOEFL), includes audio material to assess test-takers' ability to comprehend

spoken English. Similarly, PISA and NAEP have developed science items that include experiments and simulations controlled by the test-taker as part of the prompt. In both cases, the item response interaction typically requires test-takers to select an answer from a set of options. In contrast, PARCC and SBAC have focused on enhancing item response interactions. In mathematics, for example, test-takers interact directly with coordinate planes to plot points and lines. Items with response formats that deviate from common selected-response and constructed-response are considered technology-enhanced items (TEIs).

TEIs are often claimed to: (a) reduce construct irrelevant variance by improving the authenticity of the contexts presented to test-takers and (b) improve construct representation (Strain-Seymour et al., 2009). Digital innovations to the prompt and the response interaction space allow TEIs to present contexts that are more authentic for applying the assessed constructs. In this way, TEIs better represent real-life contexts in which the construct is typically applied. In addition, TEIs have potential to assess constructs that are not feasible to assess with selected-response items. As an example, consider this high school Common Core standard: "Graph functions expressed symbolically and show key features of the graph, by hand in simple cases and using technology for more complicated cases" (CCSS.MATH.CONTENT.HSF.IF.C.7; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). A TEI assessing this standard would allow test-takers to draw a graph on a coordinate plane and then identify important elements of their graph, such as vertices, maximums, or minimums. Since their inception, both PARCC and SBAC have

purposefully capitalized on the potential of TEIs to assess standards that have not been assessed by selected-response items (Crabtree, 2016).

The appeal of item innovation spurred the development of a variety of technology-enhanced item formats by several computer-based assessment programs. However, despite their increased popularity and prevalence among testing programs, there is limited empirical evidence regarding the benefits of using TEIs compared to those of selected-response items (Bryant, 2017). Gathering evidence regarding whether and how TEIs improve the measurement properties of a test beyond SR items will strengthen validity claims for assessment programs that have adopted TEIs and will inform those that are transitioning to digital platforms.

To date, only six studies have compared the psychometric properties of TEIs to those of selected-response items. These studies include three peer-reviewed studies published by Jodoin (2003), Wan and Henly (2012), and Qian et al. (2017) and three dissertations authored by Gutierrez (2009), Eberhart (2015), and Crabtree (2016). Although these six studies provide some insight into the benefits of TEIs, several shortcomings limit the practical applications of their findings. Salient limitations of this body of literature include: (a) grouping diverse innovative item formats under a single "TEI" label, (b) comparing TEIs to SRIs that assessed different constructs, and (c) the lack of a consistent definition of "technology-enhanced item" across studies. Bryant (2017) warns against making blanket statements about the worth of TEIs as a class given the diversity of TEI formats currently used in operational tests. Evidence indicates that some types of TEIs provide better utility than others (Russell & Moncaleano, 2019) and that the value of TEIs is dependent on the type of interaction required to produce a response

and the construct intended to be measured. Therefore, studies about TEIs should focus on specific TEI formats rather than "TEIs" as a whole. Moreover, when TEIs are compared to selected-response items that assess different constructs, any differences on item psychometric characteristics across item formats may be confounded with differences in content. Finally, the results of several studies are often not directly comparable as different authors adopt different definitions of TEIs, thus classifying a variety of item formats under this label.

The study described in this dissertation adds to this limited body of work by comparing TEIs to multiple-choice items while addressing some of the shortcomings of this group of studies. In particular, this study relies on two forms of a data collection instrument which have been constructed using pairs of TEIs and stem-equivalent multiple-choice items (MCIs), thus ensuring the equivalence of the assessed content across items. Additionally, all analyses were conducted separately for the two types of TEI response formats used (drag-and-drop classification and drag-and-drop rank-ordering), thus allowing results to be reported separately for each response interaction format. Finally, this work adopts a clear definition of what kind of item constitutes a "technology-enhanced item" and uses this definition consistently throughout.

Purpose of the Study

The primary purpose of this study is to develop a protocol to judge the comparative measurement value of technology-enhanced items and stem-equivalent multiple-choice (MC) items to inform item-type selection decisions. To do so the following research questions are explored:

1. How do the psychometric characteristics of commonly employed TEI drag-and-drop formats (classification and rank-ordering) compare to stem-equivalent multiple-choice items?

2. What is the relationship between the utility of TEI drag-and-drop formats (classification and rank-ordering) and their psychometric item characteristics?

3. How can TEI psychometric properties and utility ratings be combined to develop a standardized protocol to judge the comparative measurement value of TEIs relative to stem-equivalent MC items?

A data collection instrument was developed and administered to a sample of adults to investigate these research questions. This instrument assessed knowledge of middle-school mathematics and science, as well as statistics concepts typically covered in high-school and college-level introductory statistics courses. The instrument comprised stem-equivalent pairs of TEIs and multiple-choice items split across two forms. Two TEI formats were examined: drag-and-drop classification and drag-and-drop rank-ordering. The drag-and-drop format was chosen as the focus of this study because Russell and Moncaleano (2019) found it to be the most prevalent TEI response interaction in K-12 operational large-scale assessment programs worldwide. Item scores were analyzed quantitatively using both classical test theory and item response theory methodologies.

To address research question 1 (RQ1), stem-equivalent items were compared according to several psychometric characteristics, including item difficulty, item discrimination, item information, and relative efficiency. Research question 2 (RQ2) was investigated by asking a panel of educational measurement graduate students to rate the

construct fidelity and usability of the TEIs employed according to the TEI Utility Framework (Russell, 2016; Russell & Moncaleano, 2019). The resulting ratings were then compared to the psychometric characteristics estimated for RQ1 to explore how TEI construct fidelity and usability ratings were associated to the psychometric behavior of these items. To explore research question 3 (RQ3), some item characteristics estimated to address RQ1 and RQ2 were combined in a protocol designed to inform decisions regarding which item format to use based on the comparative measurement value of technology-enhanced items with respect to stem-equivalent multiple-choice counterparts. This protocol relies on four indicators to provide a judgment of comparative measurement value that informs when it is valuable to use a TEI instead of a MC format. The protocol was applied to the TEIs developed for this study to examine comparative measurement value of both TEI response interactions considered (i.e., classification and rank-ordering).

Significance of the Study

The educational assessment industry continues to develop new TEIs to improve measurement. However, there is a lack of empirical evidence regarding the benefit of replacing an SRI with a TEI. In other words, it is unclear when it is appropriate to use a TEI format (and which) instead of a selected-response format. Although test developers often make efforts to ensure that TEIs are properly aligned with the intended content and reduce construct irrelevant variance, no research is available about whether these efforts pay off by producing items that assess the intended constructs better than the SRIs that they were designed to replace. In part, this lack of research can be explained by the lack of tools to determine what "better" means when comparing TEIs to traditional item formats. Although guidelines for evaluating multiple-choice and open-ended items based

on their psychometric properties have been available for several decades, there are no comparable resources specific to evaluating the measurement value of TEIs compared to SRI counterparts.

This dissertation proposes a methodology to examine the measurement value of TEIs compared to that of multiple-choice items and provides rigorous evidence and practical results regarding the benefits of two types of the common drag-and-drop TEI format (classification and rank-ordering) compared to MC items. This is accomplished by addressing some limitations of the relevant literature. In particular, by conducting research on a single type of TEI format and employing stem-equivalent pairs of TEIs and MCIs to ensure the content-equivalence of items across test forms and response interaction spaces. Additionally, by addressing some of the methodological shortcomings of the limited body of work that has examined the psychometric properties of TEIs, this study aims to provide the field of educational measurement with a new standard to conduct research on TEIs. Finally, the protocol proposed as a result of this study will provide a clear indication of the tradeoffs involved whenever a TEI replaces an MCI.

The remainder of this dissertation is divided into four sections. Chapter 2 presents a review of the literature that informed the study and its design. The major topics reviewed include the growth of the testing industry over the last century, the introduction of computers as delivery platforms, and the development of digital innovations to item formats. In addition, this chapter includes a review of methods used to examine item quality, in particular, how items have been compared based on their response format, delivery mode, and interface. The chapter concludes with an overview of empirical studies that have compared computer-based innovative and traditional item formats and a

discussion of how the methodologies employed in these studies informed this work. Chapter 3 presents the design and analytic methods of this study. Specifically, this chapter describes the instrument development process, the use of the TEI Utility Framework to rate the construct fidelity and usability of the TEIs used, the instrument administration procedures, and the quantitative methods used for item analysis. Chapter 4 describes the results of the study. The chapter begins with an examination of sample characteristics and time spent by participants on the instrument. Then the chapter presents comparisons of TEIs and their MC counterparts based on the psychometric properties estimated, followed by findings regarding the construct fidelity and usability ratings of the TEIs. Finally, Chapter 5 identifies the most valuable characteristics for judging comparative measurement value and proposes a decision protocol that standardizes comparisons between TEIs and traditional item formats. Chapter 5 concludes with a discussion about the implications of the study for the field of educational measurement, the limitations of this study, and future steps on this line of research.

Throughout this dissertation, the term efficiency is used in several ways. The glossary in page 174 details each way in which the term efficiency is used and other common terms used in this work.

Chapter 2 - Literature Review

The purpose of the study is to develop a protocol to estimate the comparative measurement value of technology-enhanced items and inform item-type selection decisions. To accomplish this, this study estimates the psychometric properties of selected technology-enhanced item formats and examines the relationship between these properties and utility. This study is informed by two bodies of literature which are both summarized in this chapter: (a) the evolution of different item formats and (b) methods for examining item characteristics and the factors that influence these characteristics. The first section of the chapter is organized chronologically, showing the birth and subsequent growth of the testing industry in the early 20th century, the development of different paper-based traditional item formats, and the introduction of computers as delivery platforms. Because the availability of computers fostered digital innovations to item formats, this section of the chapter also presents efforts made to define and classify these innovations, as well as a discussion of their potential benefits and limitations.

The second section of the chapter focuses on the methods used to examine item quality. Informed by the decades of work on paper-based traditional items, item and test characteristics valued by test developers are discussed. Methodological approaches commonly used to compare items based on their response format, delivery mode, and interface are also presented. A summary of efforts pursued to examine the quality of innovative computer-based items is also included. The chapter concludes with a review of empirical studies that have compared computer-based innovative and traditional item formats and a discussion of how the limitations of these studies informed methodological considerations of this work.

Traditional Items

A Brief History of Testing and Traditional Item Formats

This section presents a brief history of educational testing primarily in the United States. The section begins with school testing practices and early efforts to develop standardized tests at the beginning of the 20th century. The growth and evolution of the testing industry throughout the rest of the century is described up to the introduction of computers as delivery platforms for educational assessments. This review focuses on the introduction and evolution of multiple paper-based item formats. The most common traditional item formats discussed in this section are described in detail in Appendix A.

Open-ended questions represent the "original" item format. Prior to the late 1800s, educational examinations took the form of oral recitations. These examinations were evaluated by classroom teachers and typically took the form of a conversation rather than a pre-determined set of questions common to all students (Russell, 2006). In the mid-1800s, however, a new format of examination was introduced. Seeking to monitor the quality of instruction by comparing schools' and teachers' performance, in 1845 Horace Mann capitalized on the new mass production of paper to create and administer common written examinations across Boston Public Schools (Gallagher, 2003; Odell, 1928; Russell, 2006). Not long afterwards, in Europe, Binet and Simon developed a measure of mental ability to determine the extent to which children possessed knowledge and skills corresponding to their chronological age (Binet & Simon, 1905; Gould, 1981). Binet and Simon were particularly interested in identifying children with mental abilities below their age that would benefit from specialized schooling; thus, their instrument became a tool to inform school placement decisions (French & Hale, 1990). Although the Binet-Simon test

was administered orally, it further promoted the idea of a standardized set of questions across all examinees (Binet & Simon, 1905). Three years later, Henry Goddard introduced the Binet-Simon instrument to America by implementing it at the Vineyard Training School for Feeble-Minded Girls and Boys in New Jersey to identify students whose mental development was below age level (Zenderland, 1998). The Binet-Simon scale was later adapted to the American context by Louis Terman and became known as the Stanford-Binet Intelligence Scale (Terman, 1916).

The use of "standard" test instruments containing open-response items expanded rapidly in classroom settings during this period. Through his dissertation, Fredrick Kelly realized that teachers were spending considerable amounts of time scoring written tests and found there was a high level of subjectivity in how these tests were marked by different teachers (Kelly, 1915). To address these shortcomings, Kelly published the Kansas Silent Reading Test which introduced the 4-option multiple-choice item format. Student responses to this test were quickly and objectively scored by scanning a test's page by eye (Kamenetz, 2015). Kelly's multiple-choice item format was reproduced in several contemporary intelligence scales. Most prominently, Arthur Otis, one of Terman's students, created a multiple-choice version of the Stanford-Binet instrument that could be administered in a group setting (Clarke et al., 2000). The utility of the multiple-choice item format was solidified in 1918 by its use in the Army Alpha, a battery of tests used to efficiently classify approximately 2 million army recruits into appropriate military positions during World War I (Carson, 1993; Monahan, 1998; Yerkes, 1921).

Following the success of the Army Alpha, the 1920s saw a rapid growth of standardized examinations in educational contexts (Clarke et al., 2000). First, Terman

adapted the Army Alpha for school settings as the National Intelligence Test. Later, Otis published the Stanford Achievement Test and the College Entrance Examination Board introduced the Scholastic Aptitude Test (SAT; Clarke et al., 2000). State-level tests also became common during this period, such as the New York Regent's exam and the Iowa Test of Basic Skills (Moncaleano & Russell, 2018). The rapid expansion and growth of these testing programs led to the creation of a variety of item formats (Odell, 1928). For example, in contrast to the traditional longer essay format, variations to the open-ended format included simple statement answers, single-answer or recall items, and sentence completion questions. Similarly, several new selected-response item formats were introduced such as true-false and matching items (see Appendix A for detailed descriptions). The 4-option MC item format was also modified by expanding ways of presenting answer options and introducing new ways for test-takers to mark their responses (e.g., by underlining, by circling, or by writing the numeral of the answer in a box or space; Douglass, 1926; Odell, 1928). The introduction of these alternative item formats prompted extensive research that compared these alternate formats to "traditional" open-ended questions (e.g., Kinder, 1925; Paterson, 1926; Ruch & Charles, 1928; Ruch & Stoddard, 1925 See also Kinney & Eurich, 1932, for a review of these early studies).

More item formats and new variations to existing item formats were introduced in the 1930s and 40s. Some new item formats included location and identification items, rank-ordering items, and detection and correction of errors (see Appendix A for detailed descriptions; Tiegs, 1939). Common item formats were also modified further. For example, it became common for the multiple-choice item to have fewer than or more than

four response options. The extended multiple-choice and extended true-false formats were also introduced, where multiple questions were answered simultaneously based on a common set of answer options (Hawkes et al., 1936). As the 1940s came to an end, authors began classifying item formats according to whether they were objectively scored (i.e., a single correct answer existed) or subjectively scored (i.e., multiple answers might be considered correct or partial credit was awarded); a classification scheme that is still used today (Haladyna & Rodriguez, 2013; Travers, 1950; Weitzman & McNamara, 1949).

The introduction of automatic scoring machines in the late 1930s and their increased availability in subsequent decades led large-scale testing programs to prefer objectively-scored selected-response item formats. As a result, no new item formats were introduced in the 1950s and 1960s. Instead, some item formats lost favor because they could not be scored automatically using technology available at the time (e.g., matching, location/identification, rank-ordering, and short open-ended items). Despite these formats being abandoned in large-scale testing, they continued to be used in classroom settings (Ebel, 1965, 1972; Ebel & Frisbie, 1991; Hills, 1976; Hopkins & Antes, 1979).

In the early 1970s and through the 80s and 90s, computers were slowly introduced as a platform for delivering tests. The first efforts to explore computer-based delivery were made by psychologists and the Office of Naval Research, which supported extensive research on adaptive testing delivered by computers (Moncaleano & Russell, 2018). Ultimately, the launch of the World Wide Web in 1989 cemented the central role of computers for the delivery of educational tests. The increased use of computers to deliver tests led to a new wave of innovation in item types. Before discussing these new item types, the use of traditional item types in digital assessment is discussed in detail.

Traditional Items in Digital Assessments

Fueled by increased computer access and the availability of internet, several testing programs have recently transitioned to computer-based delivery systems. Initial transitions occurred by simply replicating traditional items used on paper-based tests in a digital platform (Bennett, 1998; Parshall et al., 1996; Poggio & McJunkin, 2012). These efforts sparked a body of research that examined the equivalence of digitally presented traditional items and their paper counterparts (e.g., Bennett et al., 2008, Clariana & Wallace, 2002; Goldberg & Pedulla, 2002; Horkay et al., 2006; Sandene et al., 2005 —see also Kingston, 2009; Leeson, 2006; Wang et al., 2007, for reviews of the literature). These studies often found mode effects; that is, differences in the performance of examinees based on the mode of administration (Clariana & Wallace, 2002; Parshall, 2002). In response to these differences, statistical adjustments were developed to ensure the equivalence of test scores over time. Although mode effects presented challenges for test developers, transitioning to computer-based administration also brought several advantages, including improved efficiencies during item review, test distribution, and scoring (Moncaleano & Russell, 2018). For selected-response items, digital platforms allow automatic scoring of student responses through pre-programed scoring protocols. For open-ended responses, digital platforms allow for efficient collection of responses and distribution to scorers. Additionally, several large-scale testing programs have developed automatic essay scoring software that mimics human scoring (O'Leary, 2018).

As large-scale assessments began to transition to digital delivery systems, Bennett (1998) introduced a framework that described three generations of computer-based tests. First-generation computer-based tests are "substantively the same as those administered

on paper: they measure the same skills, use the same behavioral designs, and depend primarily on the same types of tasks" (Bennett, 1998, p. 3). During the second-generation, testing programs begin "delivering qualitatively different tests cost-effectively" (Bennett, 1998, p. 5). In particular, Bennett focuses on the ability to incorporate multimedia (sound and video) to measure traditional skills more comprehensively and assess new constructs. Some response formats also change during this generation. For example, substituting multiple-choice items for open-response items with a single correct response (e.g., a numeric response to a simple arithmetic problem) that the software scores automatically. Bennett also suggests computers will open the door to developing items with more than one plausible answer that can be scored automatically, as the system would be able to compare a student's response with several acceptable answers or evaluate whether an answer meets certain criteria. In other words, Bennett anticipates innovations to item response formats (i.e., technology-enhanced items) as a defining feature of second-generation CBTs.

During the third stage, which Bennett (1998) terms generation "R," electronic tests break from tradition and become embedded in the curriculum and the instructional process. During this generation Bennett believes "the influence of cognitive science will be strongly more evident driving course and test design" (Bennett, 1998, p. 12). This will lead tests to be "theory-based" by capitalizing on fundamental conceptions of the nature of each subject and the associated cognitive processes. This generation will include very different tasks from previous generations, for instance, replacing multimedia with virtual simulations or complex modeling environments.

Since Bennett (1998) introduced his computer-based testing generations framework, large-scale testing programs have evolved considerably. Bennet (2015) and O'Leary et al. (2018) argue that the field has entered the second generation of computer-based assessments and is beginning to move towards the third as evidenced by the use of simulations, automatic essay scoring algorithms, and an increasing diversity of innovative item formats. The following section focuses on these innovative items.

Technology-Enhanced Items

Computer-based test delivery provides test-developers an opportunity to create new digitally-based item formats that may be scored automatically. The following section presents an overview of definitions and classification schemes that have been proposed for these new item formats and concludes with a discussion of the potential benefits and limitations of these innovations.

Defining TEIs

Digital items that differ from traditional formats have been referred to by several terms over the last three decades. These terms include *new item types*, *innovative items*, *technology-enabled items*, *computer-based items*, *sophisticated tasks*, *interactive items*, and *technology-enhanced items* (Bryant, 2017). While "new item type" was the most prevalent label in the 1990s and "innovative item" in the 2000s, "technology-enhanced item" became the preferred label during the most recent decade. Two main approaches have been employed to characterize innovative item formats: (a) operational definitions and (b) classification schemes. Operational definitions present criteria that allow one to classify an item format as innovative or traditional. In contrast, classification schemes consider innovation as a continuum where item types are classified according to the degree of

innovation they provide based on multiple item characteristics. These two approaches are described in detail.

In the broadest sense, the label "technology-enhanced item" (TEI) refers to any computer-based item that differs from traditional selected-response and constructed-response items (Bryant, 2017; Parshall & Guille, 2016). For example, Parshall et al. (2010) define TEIs as items that "make use of features and functions of a computer to deliver assessments that do things not easily done in traditional paper-and-pencil assessments" (p. 215). A similar definition is used by Smarter Balanced in its Technology-Enhanced Items Guidelines (Measured Progress & Educational Testing Service, 2012), where TEIs are "computer-delivered items that include specialized interactions for response and/or accompanying response data. These include interactions/responses that are not selected-response or text-entry. TEIs may include digital media as the stimulus" (p. 9). Russell (2016) makes a distinction between enhancements to the prompt and the answer space.

> Technology-enhanced items fall into two broad categories. The first category includes items that contain media that cannot be presented on paper. These items utilize video, sound, 3D graphics, and animations as part of the stimulus and/or response options. The second category includes items that require test takers to demonstrate knowledge, skills, and abilities using response interactions that provide methods for producing responses other than selecting from a set of options or entering alphanumeric content. To distinguish the two categories, the term technology-enabled refers to the first category and technology-enhanced labels the second category. (p. 20)

In contrast to operational definitions, classification schemes are another approach used to describe computer-based innovations to item formats. Bennett (1993) proposed six main categories to classify computer-based item formats: (a) multiple choice, (b) selection/identification, (c) reordering/rearrangement, (d) completion, (e) construction, and (f) presentation. This classification scheme focuses on the degree of constraint imposed by the response interaction for a given item type. In a similar fashion Parshall et al. (1996) argue that item innovations reside on a continuum that ranges from fully constrained responses (e.g., multiple-choice) to highly open responses (e.g., an essay). Koch (1993) organized innovative items according to how different they were from traditional items using four hierarchical categories: (a) traditional items with minor modifications, (b) items that make fuller use of graphics and graphic capabilities, (c) "multidimensional" items (i.e., items that require test-takers to interact with content presented in a matrix form), and (d) situated items (online items with a high degree of real-world fidelity).

Based on some these early classification frameworks, Parshall et al. (2000) argued there were five dimensions of item innovation in digital environments: (a) assessment structure, (b) response action, (c) media inclusion, (d) level of interactivity, and (e) scoring algorithm. The assessment structure is concerned with the level of constraint of an item format. Item format constraint ranged from multiple-choice (most constrained) to constructed responses (least constrained). The response interaction dimension refers to the hardware used by test-takers to engage with the item (e.g., keyboard, mouse, touch screen). Media inclusion focuses on the extent to which media was included as part of the stem or the response space. The level of interactivity dimension focuses on the extent to

which an item responds or reacts to the responses provided by a test-taker. The level of interactivity ultimately refers to the immediate feedback test-takers see as they interact with an item, for example, by seeing the height of the bar in a histogram change as they move the cursor. Finally, the scoring algorithm dimension addresses how responses are translated into quantitative scores.

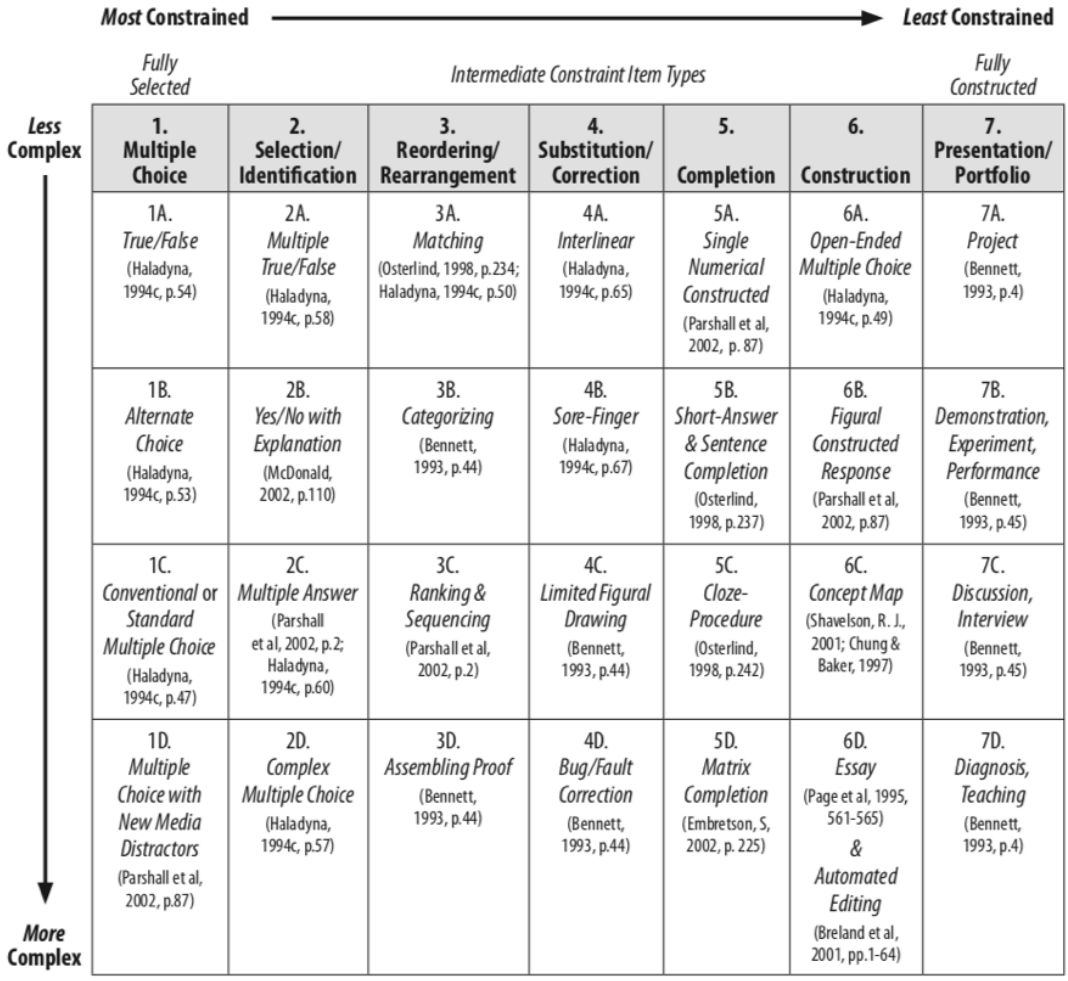
In an updated version of the chapter, Parshall et al. (2010) included two more dimensions: complexity and fidelity. Complexity relates to the "number and variety of elements examinees need to interpret and use in order to respond to an item" (Parshall et al. 2010, p. 216). Fidelity is associated with the extent to which the item provides accurate and realistic representations of objects, situations or tasks in which the construct being measured may be applied (Parshall et al., 2010). The authors further highlight that each of these dimensions may be seen as a continuum, which ranges from less to more "innovative." The authors also warn that the "maximum level" of innovation within each dimension is not always necessary or recommended; rather, test-developers should strive for the optimal innovation level of each dimension in accordance to the purpose and content of the assessment (Parshall et al., 2000; Parshall et al., 2010).

Scalise and Gifford (2006) expanded Bennet's (1993) original constraint continuum to seven levels: (a) multiple-choice, (b) selection/identification, (c) reordering/ rearrangement, (d) substitution/correction, (e) completion, (f) construction, and (g) presentation/portfolio. In addition to expanding this continuum, the authors suggested that items could vary according to their complexity within each of these constraint levels, often in the form of higher-order interactions and the inclusion of multimedia. Based on

this two-dimensional taxonomy, the authors reviewed 44 papers and book chapters and classified 28 commonly used innovative items (see Figure 2.1).

Figure 2.1

Scalise and Gifford's "intermediate constraint" taxonomy for e-learning assessment questions and tasks



Note. From "Computer-based assessment in E-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms," by Scalise, K. and Gifford, B., 2006, *The Journal of Technology, Learning, and Assessment, 4*(6), p. 9 (https://ejournals.bc.edu/index.php/jtla/article/view/1653). ©2006 The Journal of Technology, Learning, and Assessment.

The transition to computer-based tests produced a wide variety of digital platforms and methods to code, store, transfer, and present computer-based items (Russell et al., 2011). To standardize these schemes, the IMS Global Learning Consortium developed the Question and Test Interoperability (QTI) Specification (IMS Global Learning Consortium [IMS-GLC], 2020; Russell et al., 2011; Santos et al., 2012). Although the intent was not to develop a formal classification scheme, the structure used to document the various item types supported by QTI effectively serves as a classification scheme. The first version of QTI was released in 2002 and included methods for coding common test item formats such as 4- and 5-option multiple-choice as well as short- and extended-answer open response. As tests have evolved and new item types have been developed, the current QTI specification released in 2020 classifies 20 item-type interactions into four main categories: (a) simple interactions, (b) text-based interactions, (c) graphical interactions, and (d) miscellaneous interactions. Table 2.1 presents the list of interactions included in the QTI specification (IMS-GLC, 2020). Like Scalise and Gifford's classification system, the IMS scheme includes both traditional and technology-enhanced items.

Table 2.1

Computer-based item interactions and their classification under the QTI Specification

Simple	Text-Based	Graphical	Miscellaneous
Choice	Inline Choice	Hot Spot	Slider
Order	Text Entry	Graphic Order	Media
Association	Extended Text	Graphic Association	Drawing
Matching	Hot Text	Graphic Gap Matching	Upload
Gap Matching		Select Point	Portable Custom Interaction (PCI)
		Position Object	

Among these approaches to defining item innovations in digital environments, Russell's definition is the most valuable for the study proposed here because it makes a simple yet clear distinction between enhancements to the prompt and enhancements to the response space. This dissertation focuses narrowly on enhancements to the response space. Thus, henceforth Russell's label of *technology-enhanced item* and its definition are adopted, where the distinctive feature of a TEI is an innovative answer space that differs from the traditional selected- or constructed-response interactions.

TEI Formats

Increased access to computers over the last decades enabled a variety of innovations to item response spaces. In a recent survey of large-scale testing programs that use technology-enhanced items Russell and Moncaleano (2019) found seven categories of interaction types that were prevalent among computer-based tests, including: (a) drag-and-drop, (b) plotting points, (c) selecting text, (d) creating frequency plots, (e) shading areas, and (f) creating partitions. This review was limited to eight operational educational testing programs in the United States and internationally that were administered in English and had publicly available items. This study does not include other types of interactions that are often seen as common among computer-delivered tests, such as hot spot interactions, matching elements, and in-line drop-down menus. These nine TEI formats are described in detail in Appendix A.

It is worth noting, however, that there is a parallel between these TEI interactions and several of the traditional item formats discussed earlier. TEI interactions may be seen as digital adaptations of traditional items that were discarded when selected-response formats became the norm in paper-based tests due to automatic scoring. For example, the

drag-and-drop and hot-spot interactions are akin to location/identification exercises while the in-line choice and fill-in-the-blank interactions are digital versions of completion exercises. Similarly, graphing and drawing exercises were common when teachers scored tests. In this way, technology-enhanced item formats are not typically new response formats, but digital adaptations of paper-based formats that are scored in an automated manner.

Potential Benefits of TEIs

Traditional selected-response item formats have been criticized for their constrained nature and their lack of authenticity (Boyle & Hutchison, 2009; Madaus & O'Dwyer, 1999). Common criticisms include the prevalence of guessing the correct answer and the tendency of SR items to assess lower-order knowledge (i.e., simple recall and recognition; Scully, 2017). Critics argue that SR items often target constructs at the lower end of common cognitive taxonomies[1]. However, critics acknowledge that although SR items by definition are not precluded from assessing higher-order skills, it is hard to write SR items that do (Haladyna, 1997, 1999; Haladyna & Rodriguez, 2013; Martinez, 1999). Although constructed-response formats have addressed some of these concerns their administration comes with increased scoring costs and testing time (Poggio & McJunkin, 2012; Sireci & Zenisky, 2006). Some benefits of technology-enhanced items stem from the nature of computer-based testing itself and include time- and cost-efficient test delivery and response scoring (Bryant, 2017; Dolan et al., 2011; Gifford, 2017). Beyond these benefits, which are a product of any computer-based test, TEIs offer several opportunities for

[1]Cognitive taxonomies hierarchically organize the cognitive processes involved in curricular learning objectives and the assessment of these objectives. Common taxonomies include Anderson and Krathwohl (2001), Bloom et al. (1956), and Webb (1999).

improved measurement. Among these improvements are (a) reducing construct irrelevant variance by increasing the authenticity of the contexts presented to test-takers and (b) increasing construct representation (Strain-Seymour et al., 2009).

TEIs are believed to reduce construct irrelevant variance (CIV) by providing more authentic contexts for the demonstration of skills and knowledge (Association of Test Publishers [ATP] & Institute for Credentialing Excellence [ICE], 2017; Boyle & Hutchison, 2009; Bryant, 2017, Harmes & Wise, 2016; Sireci & Zenisky, 2006; Strain-Seymour et al., 2009). As the authenticity of the context of an item increases, the ability for a test-taker to demonstrate their ability on the targeted construct is less threatened by non-targeted constructs (Huff & Sireci, 2001). For example, consider the assessment of a student's ability to graph mathematical functions. SR items that target this construct often ask students to select the correct graphical representation of a function from a list of options. In contrast, a TEI version of this item requires students to create graphical representations of functions themselves, a process that is similar to how they are assessed in the classroom. TEIs have the potential of removing construct irrelevant processes associated with selecting from among a set of response options (Drasgow & Mattern, 2006), in particular guessing the correct response, as they often provide a sufficiently large number of possible responses that the probability of guessing correctly is minimal (Gifford, 2017; Huff & Sireci, 2001; Parshall & Harmes, 2014; Strain-Seymour et al., 2009). Finally, TEIs may also reduce CIV by increasing test-taker engagement (e.g., Dolan et al., 2011; Huff & Sireci, 2001; Strain-Seymour et al., 2009; Zenisky & Sireci, 2013). TEIs hold potential to create contexts with higher fidelity than those provided by SR formats and thus be more engaging for students. Although approaches to estimate

test-taker engagement have been developed for both SR items (Wise & Kong, 2009) and TEIs (Harmes & Wise, 2016), their validity is not widely accepted.

TEIs also provide an opportunity to improve construct representation by measuring constructs that have traditionally been unassessed due to the constraints of SR and CR formats (ATP & ICE, 2017; Bryant, 2017; Dolan et al., 2011; Gierl et al., 2016; Strain-Seymour et al., 2009). For example, if one believes that identifying a graph from a set of options is a different construct than creating a graph of a given function (i.e., identification vs. construction), then graphing a mathematical function is a construct that has remained unassessed by SR items. TEIs also improve construct representation as they have the potential to assess higher-order cognitive processes (ATP & ICE, 2017; Duke-Williams & King, 2001; Strain-Seymour et al., 2009; Wendt & Harmes, 2009).

Limitations of TEIs

Despite their potential benefits, TEIs have at least three challenges and limitations. These include (a) introducing CIV associated with computer literacy, (b) increased development costs, and (c) scoring challenges.

Although TEIs may reduce CIV related to the authenticity of the context, they also may introduce CIV related to the test-takers' familiarity with computers and the actions required to produce a response (Boyle & Hutchison, 2009; Parshall & Harmes, 2014; Sireci & Zenisky, 2006). Moreover, "the use of TEIs is generally driven by the functionalities offered by item authoring and test-delivery platforms, not by the constructs identified by test developers" (Bryant, 2017, p.2). This happens when test developers first choose a TEI interaction and then identify a construct to be assessed with that item format (Parshall & Becker, 2008). Allowing technology to drive measurement may be an

additional source of CIV. Parshall et al. (2010) warn that "it is important not to undertake innovations simply because they appear to be glitzy or cutting-edge. Innovation in and of itself does not ensure better measurement, nor is it equivalent to increasing validity" (p. 228).

TEIs can also "be more expensive to develop and administer, since they depend upon advanced authoring, delivery, and scoring technologies" (Bryant, 2017, p.2). A quote from a Measured Progress report speaks to this cost: "Unfortunately, TEIs have been expensive to develop and score. They have commonly been 'one-off' productions requiring custom programming, and thus were created only for large-scale assessment, where the high stakes justified the expense" (Measured Progress, 2014, as cited in Gifford, 2017, p. 6). Similarly, the PARCC Assessment Consortium estimates that developing TEIs may cost as much as five times the cost of developing traditional multiple-choice items (Russell, 2016). Although research has explored how to streamline this process through task-models and templates (Parshall & Harmes, 2007, 2014; Strain-Seymour et al., 2009), the load on item writers and software developers is significantly higher for TEIs than for SR items (Boyle & Hutchison, 2009; Parshall & Becker, 2008; Strain-Seymour et al., 2009). Furthermore, there is a lack of empirical cost-benefit studies of TEIs because information about costs and current capabilities of the standard item authoring platforms used by test developers is limited (Parshall & Harmes, 2014).

Finally, the scoring of TEIs also presents challenges. TEIs can be scored dichotomously (correct/incorrect) or, when they require test-taker decisions or offer multiple responses, they can be scored polytomously (Parshall & Becker, 2008). TEIs also offer the possibility of gathering process data, which is data that represents each action

the examinee performs as they interact with an item (Behrens et al., 2019). Examples of process data include time spent on the item, number and order of clicks employed, and identifying the elements of the item an examinee interacted with. How to harness this data for scoring purposes remains unclear. Additionally, the broad diversity of TEIs makes it challenging to draw general conclusions about TEIs as a class of items. Different TEI formats function differently and thus each format warrants independent research (Bryant, 2017). Collectively, these issues present a psychometric challenge for the use of TEIs.

Traditional Item Evaluation Procedures

The introduction of TEIs has sparked questions about whether and when to use a TEI instead of a traditional item format. To inform this decision-making process, it is informative to examine the factors that have historically influenced decisions to include an item on a given a test. In general, there are two main factors that influence decisions about the use of an item, namely technical characteristics and construct representation. As the main purpose of this dissertation is to provide a quantitative methodology to compare TEIs and traditional item formats, the following review focuses on the technical characteristics that inform item selection in the test development process.

Examining Technical Characteristics

Several item characteristics are commonly evaluated when deciding whether or not to include an item in an assessment. These characteristics include item difficulty, item discrimination, distractor quality (for selected-response items; Livingston, 2006; Schmeiser & Welch, 2006), test dimensionality, item and model fit, and item information (American Educational Research Association et al., 2014). Each of these characteristics is described in detail.

Item Difficulty. Today, item difficulty has two technical definitions, each associated with one of two primary test theory paradigms: classical test theory (CTT) and item response theory (IRT). Under CTT, difficulty is defined as the proportion of test-takers that have answered an item correctly (often known as the p-value). This definition was introduced by Thorndike et al. in 1927. Although some authors have tried to rename it as "easiness" or "facility" (Wilmut, 1975) or redefine it as the percentage of test-takers failing to answer the item correctly (Lentz et al., 1932), none of these efforts prevailed. Under the IRT framework, the probability a test-taker answers an item correctly is modeled as a function of the examinee's ability level (θ; i.e., test-takers with high-ability have a higher probability of answering the item correctly while low-ability test-takers have a lower probability of answering the item correctly). The item difficulty parameter is estimated on the same scale as ability (θ) and takes the same value as θ wherever an examinee at that level of ability has a 50/50 chance of answering the item correctly (de Ayala, 2013; Fan, 1998). Hence, a test-taker whose ability level is higher than the item difficulty has a high probability of answering the item correctly. The IRT definition of difficulty was initially proposed by Rasch in the 1960s (Rasch, 1960, 1966) and further developed in the work of Lord and Novick in 1968, and has since remained stable (Berk, 1980).

An important difference between the two test theory paradigms is that IRT is deemed to be *sample-free*, which means, that the estimation of item difficulty (and other parameters) does not depend on the sample of test-takers used. In contrast, the difficulty of an item under CTT may be different across different samples of test-takers (Haladyna & Rodriguez, 2013). There is also a practical difference between both approaches. While

IRT requires a significant number of test-takers to obtain a reliable estimate of the difficulty parameter, the CTT approach is viable for smaller sample sizes.

The difficulty of an item has been used as a criterion for item selection since its conception. Originally, Thorndike et al. (1927) argued that that an average of 50% difficulty across a test provided the best differentiation between the abilities of a group. Later, Thurstone (1932) suggested that item difficulties could range between .30 and .70 and several authors have agreed with this proposed range (Allen & Yen, 2002; Copperud, 1979). Some authors did not provide specific cutoffs for acceptable percentages but were adamant that difficulties deviating notably from the midpoint were undesirable and thus that difficulty should not be excessively large or small (Wilmut, 1975). For example, Ebel (1972) argued that tests comprised of items that were too easy or too hard reduced the variability of total scores and may waste examinees' time by asking them to answer items that were either too difficult or too easy for them. Although Thurstone's range of values was widely accepted, other authors proposed slight variations of this range. Lindeman and Merenda (1979) suggested a narrower interval (.40 to .60). Crocker (1992) argued that although the ideal range for difficulty is from .40 to .70, difficulties ranging between .20 and .90 were acceptable. Unlike the literature on CTT item difficulty, the literature on IRT-based difficulty is absent any discussions about criteria for acceptable values for the difficulty of an item. However, in practice, difficulty parameter values often range between -3 and 3 (Baker, 2001).

A note on the relation between item difficulty and other item and test characteristics. Test reliability is an estimate of the relationship between the observed score produced by a test and the true score an examinee would receive if all measurement

error was eliminated. Consequently, reliability has an inverse relationship with the measurement error inherent in observed scores (Crocker & Algina, 1986). Under the CTT framework, test-developers often strive to maximize reliability to ensure the most accurate measure possible. For this reason, proposals of ideal ranges for item difficulties presented above were guided by a desire to maximize test reliability.

Several authors (Cronbach & Warrington, 1952; Ebel, 1967, 1972; Lord, 1952; Richardson, 1936b) have shown that as the variability of item difficulties in a test increases the variability of total scores diminishes thus decreasing the reliability of the test. Therefore, tests with a large number of very easy or very hard items will be less reliable than tests containing items of moderate difficulties. Moreover, items with extreme difficulties tend to discriminate poorly between high- and low-performers, thus reducing the variance in total scores and, in turn, decreasing test reliability (Haladyna & Rodriguez, 2013; see next section: *Item Discrimination* for further details). For these reasons, most ranges described above were moderately narrow and centered around .50 because this provides a wider range of total scores (increased variance) and in turn improves test reliability.

The appropriateness of an item's difficulty also depends on the purpose of the test. Items with extreme difficulties may be desirable when assessing groups of high-ability or low-ability test-takers. For example, difficult items would be appropriate to assess student achievement in a mathematics honors course.

Item Discrimination. In the simplest of terms, item discrimination is understood as an item's ability to differentiate between high- and low-scoring test-takers

(Haladyna & Rodriguez, 2013). Despite this common understanding, the CTT and IRT paradigms employ different approaches to estimate an item's discrimination.

In the CTT paradigm, the total score is assumed to provide the best estimate of a test-taker's ability. Therefore, the ability of an item to differentiate between high- and low-performing test-takers is estimated by examining the relationship between examinees performance on the item and their performance on the test as whole. More specifically, the product-moment correlation coefficient between test-takers' item responses and total test performance serves as an estimate of the item's discrimination. Depending on how the item is scored, different versions of the correlation coefficient are used. For example, Pearson's r is appropriate if scores are assumed to be continuous, while the point-biserial or biserial correlation coefficients are used if scoring is dichotomous.

In the IRT paradigm, the probability distribution described earlier is modeled through a logistic function centered at the estimated difficulty level of the item. The item discrimination parameter is defined as the slope of this function at the point representing a .50 probability of responding correctly to the item. In other words, if a group of examinees whose ability is close to the difficulty of the item is considered, an item with a very steep slope will differentiate between examinees whose difference in ability is relatively small. In contrast, an item with a flatter slope (i.e., low discrimination) will only differentiate between examinees whose difference in ability is large. Theoretically, the discrimination parameter ranges between $-\infty$ and $+\infty$, but in practice, values often range between -2.80 and +2.80 (Baker, 2001).

In general, the CTT and IRT estimates of item discrimination are interpreted similarly: high-positive values indicate good discrimination, 0 indicates no discrimination

and negative values are undesirable. Nevertheless, more nuanced criteria for interpretation have been proposed. Under the CTT definition, Ebel suggested that the most desirable items have discrimination index values larger than .30 (Ebel, 1954). Later he proposed further divisions, indicating that items with indexes between .01 and .19 had low discrimination, moderate if between .20 and .39, and high if above .40 (Ebel, 1972). Wilmut (1975) indicates that during the 1970s, items that had discrimination index values that exceeded .15 or .20 were considered appropriately discriminating. Meanwhile, Copperud (1979) had a more stringent position, suggesting that items with difficulty levels between .30 and .70 should have a high discrimination value (above .30). In contrast, Haladyna defined items with satisfactory discrimination as those with values above .15 (Haladyna, 2004; Haladyna & Rodriguez, 2013). Within the IRT paradigm, common criteria of minimally acceptable discrimination values include 0.50 (Baker, 2001) and 0.80 (de Ayala, 2013; McBride, 1979; Urry, 1974).

Item Distractor Quality. Another characteristic often used to evaluate the quality of selected-response items focuses on the quality of the distractors included as response options. "Ideally, each of the distractors should attract some pupils, particularly those in the low group. If no one chooses a particular distractor, it may not be functioning properly" (Lindeman & Merenda, 1979, p. 114). Hills (1976) recommended four guidelines for evaluating item distractors: (a) each distractor should attract at least one examinee, (b) the correct answer should attract a higher number of high-scoring examinees than low-scoring examinees, (c) the distractors should attract a higher number of low-scoring examinees than high-scoring examinees, and (d) more than half of the examinees should choose the correct alternative. Additional statistical analyses have been proposed such as

calculating a "distractor discrimination index" (e.g., the correlation between distractor selection, i.e., chosen/not chosen, and total test score) and chi-square significance tests of expected proportions of test-takers that choose each option (Gierl et al., 2016; Haladyna & Rodriguez, 2013; Nitko & Hsu, 1984; Parshall & Becker, 2008). In the IRT framework, it is common to plot trace curves (distractor characteristic curves) to examine the probability of a test-taker selecting a distractor as a function of their ability level, thus allowing examining the attractiveness of each distractor at different ability levels.

Test Dimensionality and Reliability. A common and important assumption of measurement theory is that a set of items measure the same construct. In other words, the items and resulting scale are unidimensional. When instruments are unidimensional their scores and associated inferences make psychological and practical sense (Hattie, 1985). First identified as a desirable test property in the 1930s and 1940s, unidimensionality was closely related to the idea of homogeneity and internal consistency (e.g., Mosier, 1936; Richardson, 1936a; Zubin, 1934). For this reason, early dimensionality measures were equivalent to or based on estimates of test reliability, specifically the KR20 formula and Cronbach's Alpha. Under the CTT framework, maximizing reliability is a paramount goal; therefore, test-developers select items that optimize these estimates (often .70 is seen as the minimum acceptable level of reliability; Cortina, 1993). In particular, items with low discrimination reduce total score variability which in turn reduces reliability. Consequently items with low discrimination are often removed.

During the 1960s and 1970s principal components analysis and common factor analysis were introduced as methods to examine unidimensionality based on communalities and eigenvalues of the correlation matrix of item scores (e.g., Hase &

Goldberg, 1967; Henrysson, 1962; Neill & Jackson, 1970). Principal components analysis is a data-reduction method that seeks to account for the most variance among items through the identification of principal components (Fabrigar & Wegener, 2011). The first principal component explains the maximum variance; therefore, the variance associated with this component was used as an index of unidimensionality (Hattie, 1985). It follows that items that had relevant loadings on components other than the first were flagged for removal. In contrast, common factor analysis seeks to understand the underlying structure of the relationships between items. Factor analysis can be used in an exploratory way (data-driven), where factors are allowed to arise from the data, or in a confirmatory way (theory-driven) where a priori hypotheses on the number of factors that explain the structure between items are tested. In factor analysis, a single factor that underlies the behavior of the items under consideration is expected. Consequently, items that do not support the hypothesis of a single factor are considered for removal. Although both of these approaches often lead to the removal of similar items, theory and purpose should drive the selection of one over the other. Both principal components and factor analysis are often considered tools of the CTT framework to improve the quality of tests through item exclusion.

Item and Model Fit. The IRT paradigm relies on statistical models to represent test-takers' observed response data. The choice of model is governed by philosophical, practical, and theoretical considerations. First, it is important to decide whether the data is expected to fit the statistical model or whether the model is expected to fit the data (Boone et al., 2013). While the latter is the most common assumption of IRT models, the former approach is the basis for the Rasch model. Second, the nature of the data (i.e., the

scoring scheme) is also an important consideration. For educational achievement tests, the most prevalent kind of data is dichotomous, but it is not uncommon to see polytomously scored responses. To model dichotomous data, the Rasch and the 1-, 2- and 3-parameter logistic (1-PL, 2-PL, 3-PL) models may be used, whereas the rating scale, partial credit, and graded response models are appropriate for polytomous data. Finally, whenever the main purpose of the use of an IRT model is to represent the data as accurately as possible, it is necessary to consider the complexity of the model used (e.g., the number of parameters in dichotomous models). Models with multiple parameters tend to fit the data better than models with fewer parameters. Under the IRT paradigm item performance and model accuracy are evaluated using goodness-of-fit statistics that compare predictions from the estimated model to the observed set of responses (Orlando & Thissen, 2003). These statistics are described in detail in this section. For simplicity, the present discussion will focus on tools associated with models for dichotomous data. However, all the discussed tools have been extended to polytomous models.

Item response theory estimates the probability of a test-taker answering an item correctly as a function of their ability and a set of item parameters. The most comprehensive model for dichotomous data is the 3-PL model for which three item parameters are estimated: discrimination (a), difficulty (b), and the lower asymptote (c). The estimated logistic function is centered at the difficulty parameter, the discrimination parameter represents the slope of the tangent to the logistic curve at this point, and the lower asymptote (or pseudo-guessing parameter) establishes a lower asymptotic limit for the curve. This last parameter is of particular importance for multiple-choice items, as it represents the probability of answering the item correctly by guessing (i.e., a test-taker

has a $1/n$ chance of answering correctly a multiple-choice item with n options by guessing). The 2-PL only estimates two parameters (a and b) while fixing c to 0. Finally, the 1-PL model assumes all items discriminate equally and consequently only estimates item difficulty. Choosing the model to use (i.e., how many parameters to estimate) is often rooted on which model best represents the data.

The Rasch model is mathematically equivalent to a 1-PL model where all items are assumed to possess discrimination of 1.0; however, it is a prescriptive model, which means the data is expected to fit the model (and not the other way around). To identify (and possibly remove) ill-fitting items, the INFIT (information-weighted fit) and OUTFIT (outlier-sensitive fit) statistics compare observed responses to the responses that were expected based on the estimated model. Both statistics are based on the squared standardized residuals between the observed and expected responses. For the INFIT statistic, these squared standardized residuals are information weighted (i.e., with a weight of $p_j(1-p_j)$; see next section for details) and then summed and averaged across persons. Therefore, the INFIT mean-square statistic is a *weighted* fit statistic. In contrast, for the OUTFIT mean-square statistic, the squared standardized residuals are not weighted when averaged, thus making it an *unweighted* statistic (de Ayala, 2013). The range for INFIT and OUTFIT values extends between 0 and infinity with an expected value of 1; therefore, values close to 1 are desired. Large discrepancies between observed and expected responses will cause these statistics to be larger than 1, indicating a lack of fit. Common cutoffs for an acceptable range of values are $1 \pm 2/\sqrt{N}$ and $1 \pm 6/\sqrt{N}$ for INFIT and OUTFIT respectively, where N represents the total sample size (Smith et al., 1998).

The INFIT and OUTFIT statistics are valid goodness-of-fit statistics for the Rasch model (and the 1-PL) because the total number correct is a sufficient statistic to estimate test-taker ability (as only one parameter is estimated), allowing a direct comparison between the observed and expected responses. This is not true for the 2- and 3-PL models as test-taker ability is defined as a latent variable where the predictions pertain to the patterns of responses rather than the summed scores (Orlando & Thissen, 2003). Orlando and Thissen (2000) introduced S-χ^2, a Pearson χ^2 statistic, and S-G^2, a likelihood ratio G^2 statistic, to address this and other limitations of common item fit indices for the 2- and 3-PL models at the time (Reise, 1990). These statistics rely on computing model-predicted joint likelihood distributions for each summed score to then "calculate expected frequencies correct and incorrect for each item for each summed score" (Orlando & Thissen, 2003, p. 290). Under the hypothesis of perfect fit, S-χ^2 is approximately distributed as a chi-square distribution allowing the use of a significance test to evaluate the fit of the model for each item (Cai et al., 2011). The null hypothesis is that there are no differences between the expected and observed frequencies, thus, statistically significant results indicate a lack of fit. As the models rarely fit perfectly, moderately significant results are expected and considered acceptable. Another common item fit statistic used in 2- and 3-PL models is the LD-χ^2 statistic introduced by Chen and Thissen (1997) to evaluate the assumption that items are locally independent. This statistic is computed by "comparing the observed and expected frequencies in each of the two-way cross-tabulations between responses to each item and each of the other items and it becomes large if a pair of items indicates local dependence" (Cai et al., 2011, p. 85). The authors introduced guidelines to interpret this statistic where values less than three are considered small, values greater than 10 are large, and values in between are inconclusive (Cai et al.,

2011). Extensions of these statistics have been developed for polytomously scored items, including the graded response, partial credit, and nominal response models.

Item Information. Item information is another IRT tool used to study item behavior. The IRT paradigm attempts to estimate the location of a test-taker on the ability continuum as accurately as possible. The amount of error associated with this estimation is quantified through the standard error of estimate (SEE). "The SEE specifies the accuracy of $\hat{\theta}$ [the estimate of test-taker ability] with respect to the person location parameter, θ" (de Ayala, 2013, p. 27). Therefore, the larger the value of the SEE the less confidence there is about the parameter's value. The SEE may be used to build a confidence interval around $\hat{\theta}$ where the true parameter value (θ) is expected to lie $(1-\alpha)\%$ of the time (where α represents the allotted Type I error rate; de Ayala, 2013). The asymptotic variance error of the estimate for $\hat{\theta}$ is given by the following formula:

$$\sigma_e^2(\hat{\theta}) = \frac{1}{\sum_{j=1}^{L} \frac{[p_j']^2}{p_j(1-p_j)}} \tag{1}$$

where p_j is the probability of a person answering the item correctly, determined by the IRT model of choice, p_j' represents the first derivative of the model, and L represents the number of items in the test (de Ayala, 2013).

The SEE is a measure of the uncertainty associated with the estimate of a person's location. Alternatively, its reciprocal may be used as an indication of the certainty of the estimate. In other words, how much information is available about each test-taker's location on the ability continuum. Hence, test information reflects measurement precision at each ability level and can be visualized as the reciprocal of the uncertainty of the

estimate of a person's location. Test information is calculated as follows:

$$I(\theta) = \frac{1}{\sigma_e^2(\hat{\theta})} = \sum_{j=1}^{L} \frac{[p_j^{'}]^2}{p_j(1-p_j)} \tag{2}$$

where all terms are defined as above. It is important to note that test information is a function of ability level, therefore, a test provides different levels of information at different points of the ability continuum.

The items that comprise the test are observations of test-taker behaviors; thus, the total information of the test is equivalent to the sum of the information each item provides over the ability continuum (as all of them are assumed independent). It follows that a general formulation for the item information function is:

$$I_j(\theta) = \frac{[p_j^{'}]^2}{p_j(1-p_j)}. \tag{3}$$

As detailed in Chapter 3, the study proposed here uses a 2-PL model, for which the first derivative is:

$$p_j^{'} = a_j p_j(1-p_j) \tag{4}$$

where a_j represents the discrimination parameter for item j. Substituting this derivative in Equation 3 produces the item information formula for the 2-PL model

$$I_j(\theta) = a_j^2 p_j(1-p_j) \tag{5}$$

Note that item information is a function of the probability of a person answering the item correctly (p_j) and the item discrimination parameter (a_j). This means that an item

provides different levels of information at different ability levels. Item information reaches its maximum value when $p_j = (1 - p_j)$, that is, when $p_j = 0.50$. In other words, item information is maximum at the location where a given test-taker has a .50 chance of answering the item correctly. Therefore, an item provides maximum information (with value of $a^2 0.25$) at the location of its difficulty (b_j) and is directly proportional to its discriminating power. Consequently, item information functions are unimodal and symmetric about b_j (de Ayala, 2013).

Item information should be interpreted with respect to a specific location on the ability continuum. For example, an item could provide an adequate amount of information for an ability level that is of no consequence to the test-developer (e.g., at the low end of the ability continuum). Alternatively, item information may be of importance when accuracy around a specific location in the ability continuum is necessary (e.g., at the pass/fail threshold for credentialing tests). In the latter case, information can become a tool to flag items for removal if test-developers seek to maximize test information at specific θ locations. Items that do not provide an acceptable amount of information at any point of the ability continuum may also be flagged for removal. The concept of information in the IRT paradigm is analogous to the concept of reliability in classical test theory. The more information an item provides at an ability level the higher the reliability of the score.

TEI Quality

The previous section examined a variety of item and test characteristics that are employed to inform the selection of items for a given test. The issues considered and criteria established to guide item selection were developed with a focus on traditional item types. Widespread adoption of computer-based test administration has led to the development of

novel technology-enhanced item types. As an increasing number of testing programs adopt TEIs, questions have arisen regarding the applicability of existing criteria for informing decisions about whether or not to include a given technology-enhanced item on a test. To this end, additional approaches have recently been developed to examine the utility of technology-enhanced items. Most notable is the Technology-Enhanced Item Utility Framework.

The TEI Utility Framework

Although several large-scale testing programs have begun using TEIs operationally, these programs have raised concerns about the added cost required to author these items (Measured Progress, 2014, as cited in Gifford, 2017). In turn, some test developers have questioned whether the utility of TEIs is worth these added costs. This interest in utility motivated the development of the Technology-Enhanced Item Utility Framework (Russell 2016). The TEI Utility Framework was designed to guide evaluation of "the extent to which a technology-enhanced item is designed to provide evidence from the test taker to support claims about the test takers development of the targeted construct" (Russell & Moncaleano, 2019, p. 2). The framework is grounded by the concept of Evidence Centered Design (Mislevy et al., 2003) and considers both *measurement utility* and *content utility* (Davey & Pitoniak, 2006).

The TEI Utility Framework is based on three facets of utility: construct fidelity, usability, and accessibility. Each of these facets is evaluated on a three-level scale: low, moderate, or high. Russell (2016) provides a matrix that can be used to pool the judgments made on each facet and arrive at a holistic evaluation of the utility of a TEI.

Based on this matrix, the overall utility of an item is rated on a five-level scale: low, moderate-to-low, moderate, moderate-to-high, and high.

Construct fidelity focuses on how closely the context created by an interaction space resembles a context in which a test-taker applies the construct in an authentic or "real-world" manner. The more authentic the context created by the interaction space, the greater the fidelity, and hence the greater the utility. Construct fidelity is split in two main components:

> a) the extent to which the interaction space creates an authentic context in which the construct is applied outside of a testing situation; and b) the extent to which the methods used by the interaction space reflect the methods used to produce products in an authentic context. (Russell, 2016, p. 25)

To evaluate a TEI's construct fidelity one judges whether the context presented by the item approximates the context in which the intended construct is applied and whether the interactions required to answer are similar to those used in a real-life context.

The second facet, *usability*, is a concept borrowed from software development. The usability of a digital tool or software is "a quality attribute that assesses how easy user interfaces are to use" (Nielsen, 2012, "What - Definition of Usability" section). In the case of TEIs and the Utility Framework, usability is defined as "the intuitive functionality of an interaction space and the ease with which a novice user can produce and modify responses with minimal mouse or finger actions and/or response control selections" (Russell, 2016, p. 25). Therefore, usability focuses on the easiness of using the interaction space to produce a response that is an accurate representation of the test-taker's ability

level. In this sense, usability is a judgment of how much construct irrelevant variance is being introduced to an item by the design of the interaction space (Russell, 2016).

Finally, *accessibility* is "the extent to which the interaction space allows test-takers who are blind, have low vision, or have motor skills-related disabilities to produce a response in an efficient manner" (Russell, 2016, p. 26). This component assumes that tests and items are responsible for "accessing" the construct-related ability level of each test-taker. Consequently, it is the item's role to overcome any barriers test-takers may have to produce a response that reflects their ability level.

The TEI Utility Framework is a useful tool for comparing TEIs to one another and examining their utility. However, in this framework, utility is based solely on the evaluation of the item prompt and interaction space. Moreover, the framework depends on human judgment to evaluate each of the three components that comprise utility. Therefore, despite the detailed rubrics employed for this purpose, the resulting judgment is subjective. Finally, the TEI Utility Framework evaluates closely item design and construct validity but does not consider item psychometric properties. Hence, it may be said that Russell's framework focuses on the design of TEIs (the "front-end") without considering item properties or item behavior (the "back-end"). Although utility is an important part of the quality of a TEI, it is not the sole component. Consequently, determining the quality of a TEI based on its measurement properties remains a topic requiring additional research (Parshall & Becker, 2008).

Item Type Comparison Efforts

Just as the introduction of selected-response formats sparked a debate about choosing the best item format in the first half of the 20th century, the introduction of computer-based testing has ignited a debate about the choice between TEIs and traditional item formats. The study presented in this dissertation focuses on comparing the utility of technology-enhanced and selected-response equivalent forms. To date, very little research has examined this issue. However, a large body of research has examined similar topics. These topics fall into three broad categories: (a) studies comparing traditional item formats, (b) studies comparing the mode of test administration, and (c) studies comparing the interface design for an item. Studies in these three categories employed a variety of methodological approaches, but two main considerations stand out as relevant to this work: (a) the research design used and (b) the statistical tools employed. In the sections that follow, an overview of the issues examined in this body of work is provided highlighting key findings. This overview is followed by a summary of the various methods used to examine these issues. The section ends with a detailed review of the few studies that focus specifically on comparing technology-enhanced and selected-response items. The limitations of these studies will be noted as these limitations informed the design of the present study.

Comparisons between Traditional Item Formats

The body of work comparing selected-response and constructed-response items is extensive. Reviews of the literature show that there was a heightened interest in comparing these item formats in the decades immediately following the introduction of SR formats (i.e., the 1920s and 1930s). Although this interest waned during the middle of the century it re-emerged in the 1980s and continued through the 1990s (Hogan, 1981;

Rodriguez, 2003). These studies focused on a variety of SR item formats (e.g., true-false, matching, multiple-choice with varying number of options) and CR item formats (e.g., short-answer, completion exercises, essays). Famularo (2007) conducted an extensive review of this body of research and found that the following issues were a primary focus across multiple studies:

- reliability and validity,
- cognitive processing requirements of the two formats,
- effects of format on student-level variables such as anxiety and motivation,
- students' attitudes toward and preferences for different formats,
- effects of format on knowledge retention,
- test preparation and test performance, and
- the effect of format on item characteristics such as difficulty and discrimination.

Across the many studies that have focused on each of these issues, the findings are mixed; some studies show consistency and comparability between SR and CR items, while other do not.

In addition to the issues Famularo (2007) identified, the construct equivalence between item formats has been a focus of study. Rodriguez (2003) reviewed and summarized 67 empirical studies that examined construct equivalence between SR and CR formats. Rodriguez's meta-analysis focused on the 56 correlations reported by studies that evaluated test form equivalence. Rodriguez found that these correlations were highly heterogeneous. However, Rodriguez identified a relationship between the magnitude of the

correlations and characteristics of the test items used. In particular, he found that the mean correlation between test forms was higher (approaching unity) when the two test forms employed items with the same stem while the correlations were lower for studies that did not employ stem-equivalent items across test forms (Rodriguez, 2003).

The absence of consistent results in comparability studies between SR and CR items has led to the common understandings that no single item format is appropriate for all assessment purposes (Martinez, 1999) and that the validity of the assessment should be the primary concern when choosing an item format (Rodriguez, 2002). Haladyna (1999, 2004) argued that fidelity should be a primary consideration when choosing an item format. Fidelity is understood as the "closeness of any test task to a criterion behavior in the target domain" (Haladyna & Rodriguez, 2013, p. 43). To Haladyna and Rodriguez, the target domain is defined by the various ways in which a given behavior (e.g., performing addition problems or reading to comprehend) might occur or be applied in the real world. Thus, to be of high fidelity, the item format should elicit test-taker behavior that strongly resembles one way in which the behavior is applied outside of a testing environment. Given this goal, the choice to use an SR or a CR format should be driven by the proposed inferences about the test-taker's ability to apply the behavior in a context outside of the test based on the test-takers performance on the item. Research has shown that when SR items and CR items are written to be content equivalent (often through the use of equivalent item stems), both item formats are equally viable to support the intended inferences from an instrument (Rodriguez, 2002). When SR and CR are content-equivalent, test-developers often yield to the SR format due to its lower production and scoring costs. However, as the assessed constructs become more complex, the ability of SR

items to approximate these constructs with high fidelity diminishes and CR formats are preferred (Haladyna & Rodriguez, 2013).

Comparisons between Test Administration Mode

As computer-based test administration increased in the late 1980s and 1990s, concern about the impact test mode might have on examinee performance escalated (Leeson, 2006). This concern inspired a series of studies on the equivalence of scores between paper-based tests (PBTs) and computer-based tests (e.g., Bennett et al., 2008; Clariana & Wallace, 2002; Goldberg & Pedulla, 2002; Horkay et al., 2006; Sandene et al., 2005). In general, findings from mode-effect studies indicate that paper-based and computer-based tests that contain the same items may not produce the same results (Clariana & Wallace, 2002; Wang et al., 2007). That is, PBT and CBT scores are not inherently equivalent (Bugbee, 1996). Differences in examinee performance attributed to the mode of administration are referred to as test mode effect.

Test mode effect is common but does not always occur, and when it is present, there is no clear pattern favoring either administration mode (Mazzeo & Harvey, 1989). In a review of 23 studies conducted by Bunderson et al. (1988), 11 showed no difference, three showed scores that favored CBT, and nine showed PBT scores being higher. Kingston (2009) reviewed 81 mode-effect studies in which effect sizes were reported. Of these, 45% reported a negative effect when a test was administered in a CBT format while 55% indicated a negative effect for PBT administration. Although the majority of mode-effects were generally small (i.e., non-significant or small effect sizes; Kingston, 2009; Wang et al., 2007) the effect is not inconsequential. In particular, for high-stakes assessments, a one-point difference can result in a higher or lower performance-level

classification which, in turn, impacts decisions based on test performance (Clariana & Wallace, 2002). Overall, the literature on mode effects is not conclusive and does not allow one to predict when mode effects may occur (DePascale et al., 2016).

The test-mode effect literature is not limited to comparisons between PBTs and CBTs. Some studies have evaluated tests administered on desktop or laptop computers and concluded that student performance across these devices is relatively equivalent (Horkay et al., 2006; Ling & Bridgeman, 2013; Powers & Potenza, 1996; Sandene et al., 2005). The comparability of scores obtained from tablet-based assessments with PBT and CBT scores has also been a focus of study in recent years. Tablets allow test-takers to interact with item content using a touchscreen, an on-screen keyboard, and/or a writing stylus. These alternate methods of interacting with item content are particularly relevant when assessing reading comprehension and writing skills (Margolin et al., 2013). Analogous to studies comparing PBT and CBT scores, studies comparing tablets to other modes of administration have produced contradicting conclusions. While some authors have found a mode effect on tablets (Chen et al., 2014) others have not (Davis et al., 2015; Davis et al., 2017; Eberhart, 2015; Ling, 2016).

Comparisons between Different Item Interfaces

Several researchers have examined characteristics of the digital interface employed to deliver a test in order to explain test-mode effects (Booth, 1998). Leeson's (2006) review of the mode effect literature classifies digital interface studies according to whether they focused on (a) the presentation of digital content or (b) the interaction of the examinee with the system. The former focuses on legibility (e.g., font type, font size, screen resolution, etc.) and layout issues, while the latter examines issues specific to navigating

the interface and recording responses. This section begins with a description of research and findings associated with the legibility of digital content (which apply to all digital content), followed by a discussion of item layout (i.e., individual items vs. clusters of items). The second part of this section summarizes research that focused on the navigation of extended reading passages (e.g., scrolling) and the input mechanisms that examinees use to produce their responses (e.g., keyboards, touchscreens).

Presentation of Digital Content. Comparing the legibility of digital content to printed content is an issue that is not unique to digital assessments. Although most of the studies described below focus on the experiences of participants with digital content in general (i.e., websites and online content), many of the lessons learned apply to digital assessments. Generally speaking, these studies addressed the same question: how does reading on paper differ from reading digital media? All the referenced studies used reading comprehension tasks to evaluate the impact of different content presentation conditions. In other words, these studies explored the mode effect of stimuli and content presentation on participants' reading comprehension. Issues that have been studied that are specific to the legibility of digital content include:

- screen size and resolution (e.g., Bridgeman et al., 2003; Chen & Perie, 2018; Ziefle, 1998),
- font characteristics (e.g., Bernard, et al., 2002; Bernard et al., 2001; Bernard & Mills, 2000; Tullis et al., 1995),
- line length and number of lines (e.g., Duchnicky & Kolers, 1983; Dyson & Haselgrove, 2001; Dyson & Kipping, 1998; McMullin et al., 2002),

- interline spacing (e.g., Chaparro et al., 2004; Kolers et al., 1981; Kruk & Muter, 1984; Lay et al., 2012),
- white space (e.g., Bernard et al., 2000; Chaparro et al., 2004; Chaparro et al., 2005; McMullin et al., 2002), and
- ambient illuminance and screen brightness (e.g., Benedetto et al., 2014; Chang et al., 2013).

Evidence gathered through these studies suggests that several long-standing typographical standards (e.g., font characteristics, line length, white space) for printed media do not necessarily apply to digital content (Dillon et al., 1990; Dyson, 2004; Hartley, 1987). Leeson (2006) summarizes several key findings concerning legibility: (a) larger screens and higher screen resolution will improve readability and may reduce reading fatigue, (b) fonts of at least a 12pt. size (particularly Times New Roman, Arial, and Tahoma) and a moderate to large white space surrounding the text will improve user experience, (c) reading speed is optimized when line length falls between 74.8 and 100 characters per line, and (d) high ambient illuminance and moderate screen brightness reduce visual fatigue. Evidence also suggests that typographical standards for print media regarding interline spacing apply to digital content.

Digital assessments present particular challenges specific to the ways items and their associated stimuli are presented. Limited by the amount of information that can be included in a single screen while keeping information legible, test developers often alter the content layout of paper-based tests when adapting them to digital platforms. One alteration that is common to nearly all computer-based tests is the presentation of a single

item per screen instead of presenting several items on a single page. In an attempt to understand the effects of presenting items in groups in PBTs versus individually in CBTs, Dimock and Cormier (1991) presented verbal reasoning items in index cards to mimic the presentation of computer-based items and avoid confounding effects of computer familiarity or anxiety. Results indicated that examinees scored significantly higher on the clustered items format than on the index card format, suggesting that layout differences between grouped items and individually presented items may be part of the mode effect observed between paper-based and computer-based formats.

Interaction with Digital Content. When examining the interaction of examinees with digital assessments, two main elements are of concern: (a) how examinees navigate digital content associated with an item and (b) how examinees input their responses. Although these two issues are relevant for digital assessments in general, they are of special importance for language assessments that require test-takers to first read a passage and then produce written responses. Printed reading assessments usually present a passage and all related items on a single page; therefore, all content is simultaneously available to examinees. In contrast, in digital interfaces, the screen is often divided vertically such that the reading passage is presented on the left side of the screen and an item associated with the passage is presented on the right (Pommerich, 2004). For longer reading passages, the examinee must scroll through the text. Comparability studies in which computer-based assessments have presented all information related to an item in a single screen (e.g., a short passage and a single item) have shown small or insignificant mode effects (Bergstrom, 1992; Bridgeman et al., 2003; Choi & Tinkler, 2002; Hetter et al., 1997; Spray et al., 1989). Conversely, comparability studies in which computer-based

assessments cannot present the totality of information in a single screen have often shown larger mode effects (Bergstrom, 1992; Bridgeman et al., 2003; Choi & Tinkler, 2002; Pommerich, 2004). Some authors attribute these differences to examinees establishing visual memories of content location on printed formats. In contrast, scrolling in digital formats weakens positional cues because the spatial frame of reference changes as one scrolls through text (Kingston, 2009; Leeson, 2006; Lovelace & Southall, 1983). Overall, these findings suggest that larger mode effects are observed in tests that require navigation compared to those that do not (Pommerich, 2004).

In paper-based tests examinees mark their responses directly in the answer booklets using a pen or a pencil, whereas in computer-based tests, examinees must interact indirectly with the test through an input mechanism to provide a response. Input mechanisms are often device-specific: while computers and laptops commonly utilize a keyboard and a mouse (or a trackpad), tablets have introduced touchscreens (including digital keyboards) and digital pens (styli). Several studies have compared traditional keyboards to digital keyboards, revealing that touchscreen keyboards produce increased musculoskeletal fatigue, user frustration, and decreased typing performance, often in the form of shorter compositions (Chaparro et al., 2014; Davis et al., 2015; DePascale et al., 2016). This effect may be minimized by the use of certain external keyboards for tablets. Research indicates that keyboards that provide physical feedback (e.g., a "click" sound, or a tactile feeling of pressing a button) improve writing performance when compared to those that do not (e.g., digital keyboards, or flat keyboards embedded in tablet covers; Chaparro et al., 2013; Chaparro et al., 2014).

Given that examinees must use their fingertips to interact with content presented on a tablet touchscreen, the small size of tablets may also compromise the precision of examinee interactions (e.g., drag-and-drop) when objects are small or close to each other (DePascale et al., 2016; Eberhart, 2015; Way et al., 2016). Although features like the two-finger *pinch-to-zoom*[2] capability (available in most tablets) have helped to mitigate some of these concerns, these features add a level of complexity to the usability of the digital interface, and in turn, to the interaction of the examinee with the assessment. Additionally, the use of tablet styli has been introduced in an effort to emulate paper-and-pencil tests, increase precision, allow examinees to provide calculations, and produce extended written responses. Although promising, this approach is not easily implemented. As an example, the Third International Mathematics and Science Study (TIMSS) recently abandoned its efforts to use tablet-stylus items in eTIMSS 2019 after a pilot study revealed that the use of a stylus increased testing time and heightened examinee stress and frustration (Fishbein, 2018).

Methodological Approaches Employed

Studies that examined item format, mode, and interface effects have employed a diverse array of methodological approaches. These approaches may be characterized according to (a) the research design and (b) the statistical methods employed. Two main research design characteristics that are of direct relevance to the proposed study are discussed, namely: (a) the degree of content equivalence between items in different forms and (b) the use of within-subject or between-subject designs. In this section, the term "test forms"

[2] "Pinch-to-zoom refers to the multi-touch gesture that zooms in or out of the displayed content on a device with a touch screen. These devices include smartphones and tablets. To use pinch-to-zoom, touch two fingers on the touch screen, and move them apart to zoom in, or together to zoom out" (Computer Hope, 2019).

will be used to describe two forms of a test that differ with respect to the characteristic under study. Depending on the focus of the study, the characteristic that differs between forms may be the item type (e.g., selected-response vs. constructed-response) or mode of administration (e.g., computer-based vs. paper-based). The reviews of the literature by Rodriguez (2003)—focused on item format comparisons—and Mazzeo and Harvey (1998) and Mead and Drasgow (1993)—focused on test mode effect—will be discussed in detail to document the variety of research designs and statistical methods employed.

In his review of 67 studies, Rodriguez (2003) classified studies based on the degree of construct equivalence between item counterparts. To this end, he classified items into one of three categories: (a) stem-equivalent items, (b) content-equivalent items, and (c) non-content equivalent items. Stem-equivalent items are item pairs that share identical stems with occasional minor modifications that direct the examinee how to respond to the item (e.g., select vs. explain). Content-equivalent items are pairs of items that do not share the same stem but are designed to target the same construct, often at the same level of cognitive demand. Finally, non-content equivalent items are items that do not share a stem and do not target the same constructs (Rodriguez, 2002, 2003). Rodriguez found that the design of test items is a significant moderator of the correlation between test-forms when comparing selected-response and constructed-response formats. Of the 49 peer-reviewed articles in Rodriguez's review, 22 used stem-equivalent pairs of items, six employed content-equivalent items, and 21 non-content-equivalent items. In the test mode effect literature, however, the majority of studies employ stem-equivalent items across forms (Clariana & Wallace, 2002).

Item comparison studies may also be classified according to whether the same or different participants interacted with each test form, that is, whether studies employed a within-subjects or a between-subjects design. Within-subjects experimental designs ensure that the ability level of subjects answering a pair of items is identical because the same participants are interacting with both test forms. A drawback of this approach is that differences observed in examinee performance between test forms may be confounded by testing practice and recall, in particular for studies that use stem-equivalent items and/or which do not counter-balance the order in which forms are administered. While between-subject designs eliminate these issues, they introduce additional challenges. Specifically, between-subjects experimental designs may introduce differences between groups (e.g., prior ability or familiarity with computers) that confound efforts to isolate item-format or mode effects. For this reason, between-subjects studies often control for the overall equivalence of the groups through random assignment or stratified random assignment of test-takers to the test form or condition (Mazzeo & Harvey, 1988). Mazzeo and Harvey's (1988) review examined 38 studies published between 1961 and 1982. Twenty-four studies used within-subjects designs and the remaining 14 employed between-subjects comparisons. The authors reported that 22 within-subjects studies controlled for the form of the test (i.e., the same test was administered) and 20 of them controlled for the order in which items were administered (i.e., counterbalancing the administration of both tests). Meanwhile, 13 of the 14 between-subjects studies controlled for subject characteristics either using random assignment or ability-matched pairs of examinees. Mead and Drasgow (1993) reviewed a total of 28 studies. Five studies employed a between-subjects design while the remaining 23 administered both test forms to the same examinees. Of the peer-reviewed studies cited in Rodriguez's review, 36 employed within-subjects designs,

and the remaining 13 relied on between-subjects designs. In sum, over all the studies included in these three reviews, about three-quarters employed within-subjects designs.

There is a strong relationship between the content equivalence of the items employed across test-forms and the research designed used. The use of stem-equivalent items across test forms in within-subjects designs is a threat to the validity of the results due to practice and recall effects. This threat is often mitigated by allowing some time to pass between administrations. For example, among the studies reviewed by Rodriguez, the time between test administrations ranged from 1 day to 5 weeks (e.g., Ackerman & Smith, 1988; Hurd, 1932; Hurlbut, 1954; Rowley, 1974). This suggests that within-subjects designs may be divided further into two categories: studies that administered both forms simultaneously and those that allowed some time to pass between form administrations (repeated measures). Using non-stem equivalent items and/or non-content equivalent items across forms are additional approaches that prevent recall and practice effects on within-subjects designs.

Table 2.2 summarizes the relationship between item content equivalence and research design for the 49 peer-reviewed studies included in Rodriguez's (2003) review. As seen in Table 2.2, stem-equivalent items are most often used in conjunction with between-subjects designs and repeated measures within-subjects designs, while non-stem-equivalent items were most commonly used in within-subjects studies in a single administration. A detailed description of the designs of these studies is presented in Appendix B.

Table 2.2

Relationship between item content equivalence and research design study among peer-reviewed studies cited in Rodriguez (2003)

Study Design	Item equivalence		
	Stem	Content	None
Within-Subjects			
Repeated Administrations	7	1	3
Simultaneous Administration	3	5	17
Between Subjects			
Simultaneous Administration	12	-	1

A final methodological component that is relevant to the study presented here focuses on statistical techniques used to compare test forms. Across the studies included in Rodriguez's (2003) review the following analytic methods were employed to evaluate the equivalence of CR and SR item formats:

- *Correlations and Partial Correlations*: Correlational studies used the correlation between total test scores in both forms as a measure of the equivalence of the item formats under scrutiny (e.g., Bennett at al., 1990; Davis & Fifer, 1959; Godshalk et al., 1966; Traub & Fisher, 1977).

- *t tests and Analysis of Variance (ANOVA)*: These studies tested the differences between the means of total test scores and item scores. A null result supports the equivalence of item formats (e.g., Carcelli et al., 1980; Coulson & Silberman, 1960; Gay, 1980; Sax & Collet, 1968).

- *Factor Analysis*: This group of studies used factor analysis to evaluate whether the factorial structure of the instrument remained invariant across test forms/item

formats (e.g., Bennett et al., 1991; Bridgeman & Rock, 1993; Pollock, 1997; Thissen et al., 1994).

- *CTT and IRT*: These studies estimated item parameters and compared them across test forms. Invariant item parameters supported the hypothesis that item formats were equivalent (e.g., Bridgeman, 1992; Lukhele et al., 1994; Martinez, 1991; Oosterhof & Coats, 1984).

In addition to the methods identified by Rodriguez, differential item functioning (DIF) analysis is also used to examine mode effects. "DIF is the tendency of an item to function differently in different groups of test-takers, groups defined by something other than their proficiency in the subject of the test" (Livingston, 2006, p. 423). The main premise underlying DIF analysis is that examinees that possess the same ability level but belong to different subgroups should not perform differently in a given item. Subsamples of examinees in each group of interest are matched according to ability level and statistical analyses are performed to evaluate the significance of the observed differences to test this assumption. DIF analysis is often associated with fairness and bias reviews, for example, comparing performance based on examinees' gender, race, or socioeconomic status but the premise is applicable to test administration modes. As described above, most studies employed within-subjects designs which provides ability-matched item scores.

Statistical tools used to examine DIF include logistic regression, the Mantel-Haenszel (MH) chi-square statistic, ETS' Delta, and item-parameter drift (de Ayala, 2013; Zwick, 2012). DIF may be assessed using logistic regression by predicting item performance (i.e., correct/incorrect) based on group membership. If group membership is found to be a significant predictor, it is an indication that group membership has an

impact on item performance. The MH test is a chi-square test of association that evaluates the independence of group membership and item score conditioned for each total score possible (i.e., conditioning for ability level). ETS' Delta uses the MH statistic to create an odds ratio that compares the odds of answering an item correctly based on group membership. ETS developed what is known as the ETS Delta scale, which classifies items in three categories depending on their level of DIF: A (negligible or nonsignificant DIF), B (slight to moderate DIF), and C (moderate to large DIF). An item is classified as type A if "the MH chi-square statistic is not significant at the 5% level or the delta is smaller than 1 in absolute value" (Zwick, 2012, p. 3), in contrast, a C item is that for which the delta is "significantly greater than 1 in absolute value at the 5% level and has an absolute value of 1.5 or more" (Zwick, 2012, p. 4). Finally, item parameter drift may be used to identify DIF by estimating IRT item parameters within each of the groups of interest. If there is no DIF, item parameters are expected to be similar within a reasonable margin of error. For further reading on DIF detection see Crocker and Algina (1986), de Ayala (2013), De Beer (2004), Dorans and Holland (1992), Jodoin and Gierl (2001), Monahan et al. (2007), Zieky (2003), and Zwick (2012).

Comparisons between Traditional Item Formats and Innovative Items

The final category of item comparison studies relevant to this study is those comparing traditional item formats to innovative items. Although the validity of innovative item formats was examined when they were introduced to large-scale assessments (e.g., Bennett, Morley, & Quardt, 2000; Bennett, Morley, Quardt, & Rock, 2000; Bennett & Rock, 1995; Davey et al., 1997; Enright et al., 1998), studies that explore their psychometric properties in comparison to selected-response and other traditional formats are scant. In fact, to

date only six studies have examined the psychometric properties of TEIs. These include three studies by Jodoin (2003), Wan and Henly (2012), and Qian et al. (2017) and three dissertations authored by Gutierrez (2009), Eberhart (2015), and Crabtree (2016). These studies are described chronologically below. Before doing so, it is important to note that a general limitation found in this body of research is that the definition of "innovative item" varies between authors and in some cases does not coincide with the definition of technology-enhanced item used here. Regardless of the terminology employed by the authors, technology-enhanced will be used when appropriate and innovative otherwise.

Jodoin's (2003) study relied on data gathered from two different tests of the Microsoft Certified Systems Engineer (MCSE) Certification Program, which assess knowledge of the implementation of Microsoft Windows software. These tests included multiple-choice items and two TEI formats: drop-and-connect (DC) and create-a-tree (CT). The DC format presents a problem statement, a universe of objects, and a pool of possible connection links between the objects (Jodoin, 2003). Examinees were required to create a diagram by using objects from the universe as nodes and identifying relationships between them with the appropriate connection links. After identifying objects that belong in the diagram, examinees select pairs that are connected to create a link. Examinees then must label the link appropriately. These items were polytomously scored by awarding points for every appropriate node and link. The CT format required examinees to build a logic tree in response to a question. A tree may document the order of steps required to complete a task or represent relationships that exist between specific entities. An incomplete tree (e.g., showing high-order headlines only) is provided to the examinee on a box to the left and a pool of objects to complete the tree are located in a box to the right.

The examinee selects objects from the box to the right and moves them to their proper position in the tree schematic on the left. CT items were polytomously scored by evaluating the use of the correct objects in the pool and their correct positioning. Appendix A includes illustrations of the items included in Jodoin's article.

Jodoin (2003) fit three-parameter logistic (3-PL) IRT models to dichotomously scored items (i.e., MCIs) and logistic graded response models (GRM) for polytomously scored items (i.e., TEIs). Based on these models, Jodoin compared item difficulty and discrimination parameters, item reliability based on item information estimates, and relative efficiency across item formats. Although information functions provide an estimate of measurement error, they are limited because they do not take the ability distribution of the examinee sample into account (Jodoin, 2003). To overcome this limitation, Jodoin (2003; following Donoghue, 1994) calculated expected information, which is a weighted average of the information function that takes the ability of the pool of examinees at each point into account (details of this calculation are presented at the end of this chapter). The mean expected information provides a better measure of the information provided by each item type (Donoghue, 1994; Jodoin, 2003). The computerized delivery system also recorded the amount of time examinees spent on each item. As the distribution of item response times is often positively skewed (van der Linden, 2006; Weeks et al., 2016), the median was the preferred measure of central tendency. The ratio of expected information to median response time spent per item then provides a measure of the efficiency of each item. Relative efficiency was then defined as the ratio between the average efficiency of both item types. Findings indicated that multiple-choice items were more discriminating and less difficult on average than TEIs. However, TEIs showed higher information than

SRIs across all ability levels. This was expected given that the TEIs were polytomously scored while SRIs were dichotomously scored (Donoghue, 1994; Samejima, 1975; Thissen, 1976). Moreover, relative efficiency results revealed that SR items provided less expected information than innovative items, but also took less time to complete. Consequently, SR items provided more expected information per unit time than TEIs (Jodoin, 2003).

In her dissertation, Gutierrez (2009) compared innovative and traditional versions of a situational judgment test. The Managerial Prioritization Skills test was developed by a team of industrial/organizational psychologists at a pre-employment test development firm as part of a battery of tests suitable to predict job performance in frontline managers. Two 16-item forms of the test were developed, an innovative one and a traditional one. In the innovative version of the test, "the interface of the assessment was designed to closely mirror what frontline managers likely experience in the day-to-day jobs" (Gutierrez, 2009, p. 70). Thus, 11 items included multimedia as part of the stimuli presented to the examinee. Examinees interacted with information presented through email, voicemails, and phone calls in order to respond to multiple-choice items. The non-innovative form of the assessment included 11 stem-equivalent items to those in the first form but presented the same information without using multimedia (for example, phone calls and voicemails were presented as written transcripts). It is worth noting that Gutierrez's definition of innovative is analogous to what Russell (2016) labeled "technology-enabled" because the innovations described occurred in the stimulus and surrounding scenario rather than in the response interaction (i.e., examinees answered MC items in both test forms).

The test was administered to consumer members of the private test-development firm who were randomly assigned to one of the two forms. As a situational judgment test,

all options provided in each multiple-choice item were correct to some extent. Hence, items were scored polytomously according to the degree of correctness of each of the options as determined by a panel of subject matter experts (SMEs). After asserting the unidimensionality of both forms of the test using exploratory factor analysis (EFA), Gutierrez (2009) proceeded to fit graded response models to each of the items. Based on these models, item parameters, and item and test information functions were reported and compared. To measure the relative efficiency of innovative versus non-innovative items, Gutierrez calculated the ratio of information (innovative to non-innovative) across the full range of the ability level scale (θ) using the test-information functions. This provided a new curve of efficiency as a function of theta. Using this relative efficiency curve, Gutierrez calculated a weighted average by integrating across the ability range with normally distributed weights. The resulting average was multiplied by the ratio of the total average time it took for examinees to complete each test form (non-innovative to innovative —note that it is the opposite of the information ratio). This final value was a measure of the relative efficiency between both test forms per unit of time. Finally, Gutierrez evaluated the face validity of the items using a post-assessment Likert-type questionnaire. Results of her study indicated that innovative items provided more information at the lower end of the ability level while traditional items provided more information at higher levels of the ability level. Moreover, innovative items were shown to provide greater measurement efficiency per unit of time. Finally, participants found innovative items to provide greater on-the-job realism and be more engaging.

In their study, Wan and Henly (2012) compared multiple-choice items to two innovative item formats (constructed-response and figural response items). Their analyses

focused on reliability, efficiency, and construct validity. Figural response (FR) items require the examinee to interact with figural material such as illustrations, diagrams, and graphs. In these items, examinees produce their responses directly on the figure provided: for example, by identifying and selecting specific areas of the illustration (hot spot interaction), drawing elements onto the illustration (e.g., arrows), or dragging-and-dropping elements to pre-determined positions (e.g., labels in a diagram). Data for their study was obtained from a statewide science achievement test aligned with the state's content standards administered to fifth-grade, eighth-grade, and high school students. Wan and Henly's study builds directly on the methodology introduced by Jodoin (2003). They compared item formats by first estimating item parameters and information functions using an IRT 3-PL model. Next, they employed a ratio between a weighted item information index (expected information) and mean response time to evaluate item format efficiency. Additionally, the authors fit confirmatory factor analysis (CFA) models (one factor and three factors) to assess construct equivalence.

Based on their findings, Wan and Henly (2012) reported that FR and MC items were equally discriminating and showed higher discriminating power than CR items. They also found that grade level moderated the relationship between item format and difficulty. Moreover, the FR items provided similar information and efficiency to MC items, while CR items provided noticeably more information than MC items but at a less efficient rate. The authors also reported that MC items and the two other item formats assessed similar constructs. Although Wan and Henly classified CR items as innovative, these items do not subscribe to the definition of technology-enhanced items used in this study. Nonetheless, their results can be interpreted as a comparison between FR items and traditional item

formats (MC and CR). From this perspective, results do not provide evidence of a clear advantage (or disadvantage) in using FR items compared to traditional item formats in any of the analyzed criteria.

Eberhart (2015) sought to compare student performance on innovative and multiple-choice items delivered through computers and tablets. Participants in her study were seventh-grade students from a Midwestern state who took the annual summative Mathematics and ELA assessments in digital platforms. Participants were randomly assigned one of three content-equivalent test forms comprised of the same number of items, sections, and parts. All test forms (three mathematics and three ELA) included both innovative and multiple-choice items. The innovative item formats used included: drop-down menus, drag-and-drop, graphing, matching, ordering, selecting text, and multiple selected-response. Note that except for multiple selected-response, all of these items are categorized as technology-enhanced item formats. Eberhart applied a two-way ANOVA to evaluate the impact of different item types (multiple-choice and innovative) and delivery systems (computers and tablets). Findings indicated that a significant interaction between factors was present in four of the forms (three mathematics and one ELA). For these forms, item type main effects were statistically significant with moderately large effect sizes. For the two ELA forms that did not present significant interactions, the main effect of item type was found to be statistically significant with large effect sizes. Moreover, in line with the mode effect studies described previously, all device type main effects and simple main effects were found to be statistically significant but with small effect sizes.

In her dissertation, Crabtree (2016) explored how the inclusion of innovative items impacted the construct validity of the Iowa End-of-Course Algebra I (IEOC-A)

assessment. Twelve test forms were randomly assigned to participating students for the 2012 administration of the IEOC-A. Each test form included five innovative items appended to a common set of 30 MC items. Innovative formats used in the IEOC-A included graphical modeling, drag-and-drop, matching, point-and-click, input text, and input number. Note that only the first three are technology-enhanced formats. Crabtree examined the constructs assessed by the test, the psychometric properties of the innovative items and the influence of these item properties on test characteristics. The author fit a 2-factor CFA model to test whether a single unidimensional construct was being assessed by the technology-enhanced version of the IEOC-A. The results confirmed a 2-factor structure with MC and innovative items loading on different factors, thus suggesting that innovative items added a new dimension to the test. A 3-PL IRT model was used to concurrently calibrate all items used in the assessment. Based on this calibration, all MC and innovative items were reorganized into three test forms that matched the content specifications of the original MC test and were equally difficult. These four forms (the original MC form and three that included innovative items) were used for further analysis. Item and test information functions were estimated for each of these four forms and compared, using the MC form as the reference point. Finally, relative efficiency was calculated as the ratio of the information functions of the three innovative forms to the MC form. Results indicated that innovative items provided more information and measured the construct more efficiently at higher ability levels than MC items (Crabtree, 2016).

Finally, Qian et al. (2017) addressed three research questions: "1) Are innovative items more psychometrically sound? 2) Do test-takers take more time to answer innovative

items? 3) Do innovative items assess higher-order thinking skills?" (Qian et al., 2017, p. 98). To answer these questions, the authors used data from an operational administration of the National Council Licensure Examination for Registered Nurses (NCLEX-RN). Seven types of items were included in this exam: simple text-based MC, MC with graphics as options, MC with an exhibit, MC with audio, multiple-response, ordered-response, and fill-in-the-blank calculation. The authors classified these items according to two of the seven innovation dimensions included in Parshall et al.'s (2010) taxonomy: assessment structure and fidelity. They identified four item formats under the assessment structure dimension: (a) multiple-choice (MC), (b) multiple-response (MR), (c) ordered-response (OR), and (d) fill-in-the-blank calculation (FC). They also categorized items in five groups under the fidelity dimension: (a) text-based items, (b) items with audio, (c) items with graphics, (d) items with an exhibit, and (e) items with graphics and an exhibit. To address their first research question, the authors compared items according to difficulty, discrimination, guessing parameters, and information functions estimated using a 3-PL model. To examine the second question, authors relied on timing data obtained from a pilot study of these items rather than their operational administration (a limitation to their study). To answer their third question, a panel of SMEs coded the items according to four cognitive levels: (a) knowledge, (b) comprehension, (c) application, or (d) analysis.

Similar to Wan and Henly's study (2012), the definition of innovative item used by Qian et al. (2017) does not correspond with the TEI definition used in this study. Of the four item formats considered in Qian et al.'s study, only ordered-response items qualify as TEIs. For these reasons, the present summary of their findings focuses on the results of the assessment structure dimension, with a particular focus on the OR item format.

Regarding research question 1, results indicated that OR items are significantly harder than FC and MC items but significantly easier than MR items. The mean discrimination parameter for OR items ranked second among the four item types considered and was not found to be significantly different from any of the other three. Similarly, the mean guessing parameter for the OR items ranked second. FC items were found to be statistically significantly harder to guess compared to MC, MR, and OR items, while no statistically significant differences were found between OR items and the remaining two formats. Upon examining information functions, the authors found that for low-ability examinees ($\theta < 1$), FC items provide more information than all other items types, while for high-ability examinees ($\theta > 1$), MR items provided the most information. While their ranking in the lower end of the ability continuum is indistinguishable, OR items ranked second at the higher end of the continuum, indicating this format provides more information than MC and FC items for high-ability examinees. Additionally, the authors grouped all innovative items together and reported that they provide more information than MC items at all levels of the ability continuum. Concerning research question 2, authors found that OR items require significantly less time to complete compared to FC items, but significantly more time than MR and MC items. Finally, regarding the third research question, there was no evidence that OR, MR, or FC items assessed higher-order cognitive skills more consistently than MC items.

Limitations of studies comparing traditional item formats and innovative items. In addition to the obvious limitation that only six studies have compared innovative items to traditional item formats shortcomings of these studies leave several questions unanswered. In particular, cross-cutting limitations of these studies

include: (a) the nature of the instruments used and the equivalence between forms, (b) the definition of an innovative item, and (c) the practical applications of the results. To aid this discussion, Table 2.3 provides a summary of the main characteristics of the six studies described above. In addition to highlighting general limitations of this pool of studies, this summary also provides an overview of methodological similarities between these studies. This section ends with a discussion of how the limitations and methodological similarities between these studies inform this dissertation.

All six studies were secondary data analyses performed on data collected from operational administrations of large-scale assessments. In all studies except for Gutierrez's (2009), the innovative and multiple-choice items weren't content-equivalent but rather targeted different constructs of the tests' target domains. The use of non-equivalent items is a threat to the validity the studies' findings because differences observed between item formats are confounded with differences in the constructs assessed. Although items in Gutierrez's study were content-equivalent across forms the differential feature was the inclusion of media in the stimulus rather than a different response interaction. In sum, there are no studies that compare TEIs to stem-equivalent—and thus content-equivalent—multiple-choice items.

Table 2.3

Summary of studies that compared innovative items to traditional item formats

Study	Content	Level	Item Content Equivalence	Comparisons and Item Types	Methodology	Criteria[a]
Jodoin (2003)	Microsoft Software	Adults	None	MC vs. TEIs (drop-and-connect*, create-a-tree*)	IRT – 3-PL IRT – GRM	Item Parameters Item Information Expected Information Relative Efficiency
Gutierrez (2009)	Managerial Skills	Adults	Stems	Text-based stimulus vs. Multimedia stimulus (all items were MC)	IRT – GRM	Item Parameters Item Information Relative Efficiency Face Validity Survey
Wan & Henly (2012)	Science State Test	K-12	None	MC vs. Innovative (figural response*, constructed response)	IRT – 3-PL	Item Parameters Expected Information Relative Efficiency
Eberhart (2015)	Math and ELA State Test	K-12	None	MC vs. Innovative (drop-down menus*, drag-and-drop*, graphing*, matching*, ordering*, selecting text*, and multiple-selected response)	ANOVA	Total Score Differences
Crabtree (2016)	Algebra 1	K-12	None	MC vs. Innovative (graphical modeling*, drag-and-drop*, matching*, input text, input number, and point-and-click)	IRT – 3-PL	Item Difficulty Item Information Test Information Relative Efficiency
Qian et al. (2017)	Nursing Licensure	Adults	None	MC vs. Innovative (ordered response*, fill-in-the-blank calculation, multiple response)	Rasch IRT – 3-PL ANOVA	Item Parameters Time Cognitive Skill Coding

*Item formats that fulfill the definition of technology-enhanced items used in this work. [a]Jodoin (2003) and Wan and Henly (2012) share the same definition of relative efficiency, Crabtree (2016) and Gutierrez (2019) use different definitions.

The second limitation of these studies is the varying definition of an innovative item. For example, Gutierrez (2009) considered items with multimedia embedded in the prompt as innovative despite all of the items relying on selected-response formats. In contrast, Wan and Henly (2012) classified constructed response items (both short- and extended-response) and figural response items as innovative, while Qian et al. (2017) used this label to group ordered-response, fill-in-the-blank calculation, and selected multiple-response items. The diversity of uses of the term "innovative" suggests authors tend to use it to describe any item format that differs from a text-based multiple-choice item. Given the lack of cohesive use of this term across the literature, it is not possible to draw overall conclusions about how the use of innovative items compares to traditional items types. It also demonstrates why a clear definition of technology-enhanced items is needed and has been explicitly adopted for this dissertation.

Applying the definition of TEI adopted for this study to these six studies it is evident that Gutierrez's (2009) study did not employ TEIs at all, and the studies by Wan and Henly (2012) and Qian et al. (2017) only included one TEI format each (the figural response item and the ordered-response item respectively). Although Eberhart (2015) and Crabtree (2016) use the label "technology-enhanced item" in their dissertations, their work includes some item formats that do not subscribe to the definition of TEI adopted in this dissertation (multiple selected-response items in the former and alphanumeric text input and point-and-click items in the latter).

The third limitation of this group of studies is the lack of useful and actionable results. Results from the cited studies have limited practical value because either (a) the study examined item formats that are not common or (b) the study grouped multiple TEI

formats for analytical and reporting purposes. The two TEI formats employed in Jodoin's study (2003), drop-and-connect and create-trees, are not commonly used in large-scale testing programs, hence the conclusions have limited application to today's educational testing programs. The figural response item format discussed in Wan and Henly (2012) is a category that encompasses multiple TEI response interactions simultaneously (e.g., hot spot, drag-and-drop, plotting points). Similarly, Eberhart (2015) and Crabtree (2016) studied instruments that used multiple TEIs that are common in large-scale assessments but grouped all TEI formats together in their analyses. The grouping of different TEI formats in these three studies precludes drawing practical conclusions for the development and use of specific technology-enhanced item formats. Given the diversity of TEI formats, Bryant (2017) advises against blanket statements about TEIs as a class, and instead advocates for studies that focus on specific TEI formats. Only Qian et al.'s (2017) work provides practical results about the benefits of a specific TEI format, namely ordered-response items.

There are several methodological similarities between these studies. Five of the six studies employed IRT models (3-PL for dichotomous data and GRM for polytomous data) to estimate item parameters as well as item and test information functions. Eberhart's (2015) study was the exception, using an ANOVA to compare total scores based on item format and delivery platform. All studies that used IRT models compared item parameters of difficulty and discrimination across item formats. Although most of the studies simply compared parameter values without performing any test to estimate the significance of differences between the values, Qian et al. (2017) used ANOVAs to evaluate the significance of the observed differences. Moreover, these five studies compared items either using item information functions or test-information functions. Some studies

reported average item information curves discriminated by item format (Qian et al., 2017; Wan and Henly, 2012), while others grouped all "innovative" items together according to their respective definitions (Crabtree, 2016; Gutierrez, 2009; Jodoin, 2003).

Finally, four of the five studies that employed IRT models also attempted to use a measure of relative efficiency of the innovative item formats to multiple-choice items. The definitions of relative efficiency were diverse. The simplest definition of relative efficiency, introduced by Lord (1980), consists in calculating the ratio of the information functions associated with two test scores at each point of the ability scale range. Therefore, the relative efficiency of test score y with respect to test score x is characterized by the following equation:

$$RE\{y, x\} = \frac{I\{\theta, y\}}{I\{\theta, x\}}. \tag{6}$$

"Scores x and y may be scores on two different tests of the same ability θ, or x and y may result from scoring the same test in two different ways" (Lord, 1980, p. 83). Plotting this function produces an easily interpretable curve of relative efficiency as a function of ability. Whenever the resulting curve is equal to 1, it suggests both formats are equally efficient at that ability level. Crabtree (2016) compared innovative items to multiple-choice items using Lord's methodology by constructing three innovative item forms and a fourth test form comprising multiple-choice items. Gutierrez's (2009) approach used Lord's function to calculate a normally weighted average of relative efficiency of two test forms. The product of this average and the ratio of the total average time for each form provided a measure of relative efficiency proportional to time.

In contrast, Jodoin (2003) calculated relative efficiency between technology-enhanced items and MC items as the ratio of average efficiency of both item formats

following a five-step process. First, information functions were calculated for each item. Second, information functions were summarized as point-estimates by calculating *expected information*. Jodoin (2003; following Donoghue, 1994) argued that although information functions provide an estimate of measurement error at all proficiency levels, they do not provide a measure of the congruence with the ability distribution of the examinee pool. For this reason, the author recommends using expected information which weighs information at each point in the ability scale using the ability distribution of the sample:

$$E(I_j) = \sum_{q=1}^{Q} I_j(\theta_q) w_q \tag{7}$$

where $I_j(\theta_q)$ is the information for item j at quadrature point q and w_q is the weight of the posterior ability distribution associated with the quadrature point[3]. Third, the measurement efficiency of each item was calculated as the ratio of average expected information to median response time. Fourth, for each item format (innovative and multiple-choice) measurement efficiency was averaged across items. Finally, relative efficiency was calculated as the ratio between the average measurement efficiency of each item formats. Wan and Henly (2012) employed the same definition as Jodoin, but used mean response time rather than median response time to calculate item efficiency. Differences in definitions of relative efficiency prevents one from directly comparing results from different studies based on this criterion. Additionally, given that all of these studies grouped both TEIs and traditional items under an "innovative item" label, the comparisons of their relative efficiency do not provide actionable results.

[3]Note that this formula is a numeric approximation of the weighted area under the item information curve $I_j(\theta)$ with the ability distribution of targeted examinees $f(\theta)$ as a weight. $E(I_j(\theta)) = \int I_j(\theta) f(\theta) d\theta$.

This dissertation builds on the methodological common ground of this body of literature while addressing several of the limitations described above. In particular, in this study, comparisons across TEIs and multiple-choice items are based on stem-equivalent item pairs and results are separated by TEI format to provide clear and practical results regarding each of the TEI response interactions used. Finally, this dissertation uses some of the item characteristics discussed in this chapter to propose a protocol that compares TEIs to MCIs in a consistent and replicable manner.

Summary of the Literature

The literature review presented in this chapter was organized in two parts. In the first part of the chapter, the evolution of traditional and technology-enhanced items was described in detail, including an overview of the transition to computer-based testing, definitions of technology-enhanced items and a discussion about their potential benefits and limitations. The second part of the chapter examined several efforts made to evaluate the quality of TEIs. Methods and criteria that have been used to evaluate the quality of traditional item include examination of: item difficulty and discrimination, item distractor quality, test dimensionality, reliability, item and model fit, and item information. Moreover, broad bodies of work that have compared items based on their format, delivery mode, and interface were described to highlight methodological considerations relevant to this work. However, the extensive literature available has relied mostly on traditional item formats, begging the question of whether the criteria and methods employed are pertinent to technology-enhanced items or whether new approaches are warranted. In the words of Parshall and Harmes (2014) "the multiple-choice item, for example, was developed and

refined over many years of use and a variety of exam programs. That deep broad level of knowledge and understanding is not yet present for innovative items" (p.16).

Although Russell (2016) introduced the TEI Utility Framework as an instrument to evaluate specifically the utility of TEIs based on their construct fidelity, usability, and accessibility, this framework is based on human judgment and relies on the design and delivery of TEIs without any regard to psychometric item properties. This limitation compounds with a general gap in the literature regarding the psychometric properties of TEIs (Bryant, 2017; Crabtree, 2016; Qian et al., 2017; Wan & Henly, 2012). Only six studies have attempted to address this deficiency and several limitations of this body of work prevent these studies from providing practical results for test and item development. These limitations include variations in the definition of "innovative" and "technology-enhanced" items, items in different formats assessing different constructs, and the overall tendency to report results at the aggregate level.

This dissertation draws on the methodological similarities of this limited body of work to compare common technology-enhanced item formats and selected-response items based on characteristics that have been traditionally used to assess the quality of traditional items. Moreover, the study design and the reporting of the results attempt to address the limitations found in previous studies. Ultimately, this study informs the proposal of a protocol to standardize comparisons between TEIs and traditional item formats and inform item-type selection decisions.

Chapter 3 - Methodology

The primary purpose of this study is to develop a protocol to evaluate the comparative measurement value of technology-enhanced items and stem-equivalent multiple-choice items. The intent of this judgment is to inform a decision as to whether to use the TEI or multiple-choice item format. To accomplish this, the following research questions are examined:

1. How do the psychometric characteristics of commonly employed TEI drag-and-drop formats (classification and rank-ordering) compare to stem-equivalent multiple-choice items? (RQ1)

2. What is the relationship between the utility of TEI drag-and-drop formats (classification and rank-ordering) and their psychometric item characteristics? (RQ2)

3. How can TEI psychometric properties and utility ratings be combined to develop a standardized protocol to judge the comparative measurement value of TEIs relative to stem-equivalent MC items? (RQ3)

This chapter describes the methodology employed to answer these research questions, including: (a) the research design, (b) the instrument and item development process, (c) the data collection procedures, and (d) the data analyses employed.

Overview

To investigate RQ1 and RQ2, stem-equivalent pairs of TEIs and multiple-choice items (MCIs) were developed and split across two forms of a data collection instrument. This

instrument was administered to a sample of adults and responses were analyzed using a 2-parameter logistic item response theory model. RQ1 was addressed by comparing several psychometric characteristics of TEIs to their MC counterparts. The TEI Utility Framework was used to investigate RQ2 by gathering judgments of the construct fidelity and usability of the TEIs employed from a panel of educational measurement graduate students. Finally, research question 3 explores approaches that combine the psychometric characteristics estimated to address RQ1 and the utility ratings used in RQ2 to develop a protocol to judge comparative measurement value. This protocol was applied separately to each TEI developed in this study and comparative measurement value judgments were examined between response interactions (classification or rank-ordering) to evaluate whether there was any clear pattern of comparative measurement value for these TEI formats.

Research Design

Instrument Development

Two 18-item forms of a single data collection instrument were developed for the study. Both forms comprised three item blocks each containing six items. The first item block for both forms was identical to facilitate IRT calibration. Each form also contained one block of technology-enhanced items and one block of multiple-choice items. Across forms the items were stem-equivalent but the method for producing a response differed. For the MC version of each item, test-takers were asked to select the best option. For the TEI version, test-takers were asked to use a drag-and-drop interaction to produce a response. Further, half of the TEIs asked test-takers to classify objects by dragging-and-dropping their responses in labeled boxes. The second half of TEIs asked test-takers to drag-and-drop

objects of a vertical list into the correct order. Table 3.1 shows the design for each form of the instrument.

Table 3.1

Instrument form construction

Form A		Form B	
Block	Item	Item	Block
Common	CL1 CL2 RO1 RO2 MC1 MC2	CL1 CL2 RO1 RO2 MC1 MC2	Common
TEI-1	CL3 CL4 CL5 RO3 RO4 RO5	MCCL3 MCCL4 MCCL5 MCRO3 MCRO4 MCRO5	MCI-1
MCI-2	MCCL6 MCCL7 MCCL8 MCRO6 MCRO7 MCRO8	CL6 CL7 CL8 RO6 RO7 RO8	TEI-2

Note. Items in block TEI-1 are stem-equivalent to items in block MCI-1 and items in block TEI-2 are stem-equivalent to items in block MCI-2. For example, items CL3 and MCCL3 are stem-equivalent.

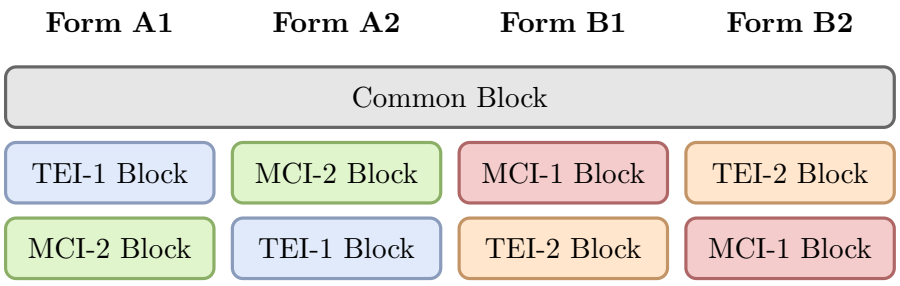
As seen in Table 3.1, the first six items (the common block) are identical across the two forms and contain two drag-and-drop classification items (CL1-2), two drag-and-drop rank-ordering items (RO1-2), and two multiple-choice items (MC1-2). The second block for both forms comprised stem-equivalent items assessing the same constructs but differing in response formats. While the second block in Form A included three classification items (CL3-5) and three rank-ordering items (RO3-5), the second block in

Form B included six stem-equivalent multiple-choice items (MCCL3-5 and MCRO3-5). Finally, the third block in Form A included six multiple-choice items (MCCL6-8 and MCRO6-8) and the stem-equivalent block in Form B included three classification items (CL6-8) and three rank-ordering items (RO6-8). Ultimately, both forms had 18 items: 10 TEIs (four of them common to all participants) and eight MCIs (two of them common to all participants). It was expected that participants would require 30 minutes or less to complete all items that formed the instrument.

At the time of administration, the order of the TEI and MCI blocks for each form was randomized to account for ordering effects. This process effectively produced four forms of the instrument (see Figure 3.1), forms A1 and B2 presented TEIs before MCIs, while forms A2 and B1 did the opposite. Additionally, forms A1 and B1 presented the content in the same order, the same was true for forms A2 and B2.

Figure 3.1

Instrument form block order



To populate the items for each form, a four-step process was followed. First, six items (two MCIs, two drag-and-drop classification items, and two drag-and-drop rank-ordering items) were developed and used to populate the common block. Second, 12 TEIs

(six classification items and six rank-ordering items) were developed. Third, these 12 items were organized in two blocks (TEI-1 and TEI-2) with three classification items and three rank-ordering items each. Finally, a stem-equivalent MCI was formed for each item included in blocks TEI-1 and TEI-2 to populate blocks MCI-1 and MCI-2 respectively. All multiple-choice items were written to have six options, a correct response, and five distractors. Items addressed concepts of high-school and college-level statistics, including: measures of central tendency, properties of frequency distributions, and interpretation of graphical data displays. Some items were adapted based on released items from the 2019 10th grade Massachusetts Comprehensive Assessment System (MCAS; Massachusetts Department of Elementary and Secondary Education [MA-DESE], n.d.-a, n.d.-b) and SAT exam (College Board, n.d.), as well as the GRE (ETS, 2009) and associated released preparation materials (ETS, 2017).

Prior to the construction of the instrument forms, two rounds of item quality review were conducted by content experts. First, all items were examined individually to ensure the prompts were clear and that how to produce a response was intuitive. Second, TEI-MCI pairs were examined to confirm their content equivalence. Although stem-equivalent items tend to be content equivalent (Rodriguez, 2003), this review ensured that each TEI and its MC stem-equivalent counterpart addressed the same construct. Based on these reviews, items were edited or, on occasion, removed from the item bank prompting the development of new items.

Participants

The target population of the proposed study was adults. As will be described later in this chapter, several psychometric item characteristics of interest were estimated using a 2-

parameter logistic (2-PL) item response theory model. The reliable estimation of this model based on participant responses was the primary concern that informed the target sample size. In the literature there is a wide variety of recommendations regarding the sample size required to fit a 2-PL model for instrument development purposes. For example, Crocker and Algina (1986) suggest a minimum of 200 participants for an item analysis study. Sahin and Anil (2017) suggest that in order to fit a 2-PL IRT model to a 20-item test, the ideal sample size is 500. Based on these recommendations, recruitment efforts were made to achieve total sample of between 600 and 800 participants. This would ensure that at least 300 responses were recorded per item.

Data Collection

Instrument Administration and Participant Recruitment

The instrument was delivered using the Qualtrics web-based survey-delivery platform. This online platform was chosen because it supports both drag-and-drop formats considered in this study (classification and rank-ordering), it allows easy test delivery to participants, and it records the amount of time respondents spend on each item. Upon completing the common item block, each participant was randomly assigned one of the four possible TEI-MCI block combinations.

Participants were recruited through Amazon's Mechanical Turk (MTurk), an online marketplace which has gained traction in recent years as a medium for researchers to gather large samples of participants for digital tasks (Barger et al., 2011; Berinsky, 2012; Cheung et al., 2017). In this platform, individuals (i.e., requesters) may post jobs in which they require worker participation. Each job, or Human Intelligence Task (HIT), represents a single assignment on which workers can work, submit a response, and receive

compensation. Workers browse available HITs and participate in them for a monetary reward set by the requester. MTurk was chosen as the best medium for instrument distribution as research has shown that data collected through MTurk for empirical studies is externally valid, reliable, and generalizable (Barger et al., 2011; Berinsky, 2012; Buhrmester et al., 2011; Cheung et al., 2017; Paolacci & Chandler, 2014; Rouse, 2015). Moreover, data gathered through MTurk has been shown to be of equal or better quality than data gathered through convenience sampling designs. Quality of the data gathered through MTurk may be ensured by refusing payment for low quality responses (e.g., random responses) or only allowing workers that have been rated highly within the system as dependable to participate (Barger et al., 2011; Buhrmester et al., 2011).

Pilot

Two hundred participants were sought through MTurk for the pilot. To qualify to participate in this study a worker had to: (a) have at least 50 HITs previously completed and approved by requesters, (b) have a 95% or higher approval rating for previously completed HITs, (c) have a high school degree, and (d) have acceptable level of English language proficiency. The first three qualifying criteria were ensured by MTurk screening procedures, while the remaining criterion was verified through a self-reporting questionnaire included at the beginning of the instrument.

Each participant responded to 20 items. Two attention-control questions were added to the 18 items developed for the study to identify careless responses. Both attention control items were based on the same stimulus and required participants to provide a simple open-ended response that did not require any calculations or critical thinking. Pilot responses were screened for quality. Participant responses were removed

for any of the following reasons: (a) participants answered incorrectly to at least one attention control item, (b) participants spent less than 5 minutes on the task, (c) participant provided incoherent answers to open-ended items, (d) responses appeared automatic. At the end of the instrument an open-ended question was included asking participants to share feedback on any of the questions, specifically whether there were any items which were unclear.

Items were revised based on the results from the pilot. Revisions were inspired by feedback provided by the participants and a simple review of classical test theory item characteristics. In particular, excluding the attention-control items, items were flagged for replacement if they were extremely easy (more than 83% correct responses) or extremely difficult (less than 16% correct responses). These cut-offs were determined by the chance of guessing a correct response in a 6-option multiple choice item. Items that showed a percent of correct responses between 16% and 30% and between 70% and 83% were also flagged for possible revision.

As will be described in Chapter 4, results of the pilot prompted the revision or removal of multiple items. Consequently, the content of the data collection instrument was broadened to include middle school mathematics and science items adapted from 2019 MCAS released items (MA-DESE, n.d.-a, n.d.-b). Additionally, a second pilot was conducted to evaluate the performance of the new items.

Operational Administration

After the items were revised following the pilot, the instrument was finalized and prepared for the final operational administration. For this stage, 800 participants were recruited through MTurk. Qualifications to participate in this study were the same as the pilot,

that is, a worker had to: (a) have at least 50 HITs previously completed and approved by requesters, (b) have a 95% or higher approval rating for previously completed HITs, (c) have a high school degree, and (d) have acceptable level of English language proficiency. The first three qualifying criteria were checked based on statistics reported by MTurk, while the remaining criterion were verified through a self-reporting questionnaire included at the beginning of the instrument.

Each participant was presented with 20 items, 18 items were developed for the study and two attention-control items intended to ensure participants were attentive while working on the task. Participant responses were removed for any of the following reasons: (a) participants answered incorrectly to at least one attention control item, (b) participants spent less than 5 minutes on the task, (c) participant provided incoherent answers to open-ended items, or (d) responses appeared automatic. Whenever participants' responses were removed, responses to all items were removed.

TEI Utility Ratings

TEIs were evaluated using Russell's TEI Utility Framework (2016) to determine the relationship between psychometric item characteristics of the TEIs included in the assessment and their utility (RQ2). The TEI Utility Framework considers three facets of utility: construct fidelity, usability, and accessibility. Construct fidelity focuses on how closely the context presented to a test-taker in a TEI resembles a context in which the construct may be applied in an authentic manner. Usability focuses on the intuitiveness with which a novice user can produce and modify a response in the interaction space of a TEI. Finally, accessibility is the extent to which the interaction space provided by a TEI allows test-takers with limitations that impact their ability to produce a response in an

efficient manner (e.g., test-takers who are blind, have low vision, or have motor skills-related disabilities). Each of the utility facets is evaluated on a three-level scale: low, moderate, or high.

While construct fidelity varies from item to item depending on the design of the item and its response interaction space, usability and accessibility depend on the test delivery system itself and how it implements the response interactions under consideration. Because Qualtrics, the instrument delivery platform used in this study, does not provide any accommodations for respondents that are blind, have low vision, or have motor skills-related disabilities, rating the accessibility of the TEIs included in the instrument was deemed uninformative. Hence, this study focuses solely on the construct fidelity and usability components of the TEI Utility Framework. Guidelines to rate TEIs on these two components are described next.

Construct fidelity considers both the context provided by the TEI and the actions required of the test-taker to produce a response. A TEI is deemed to have high construct fidelity when the "context created by the interaction space authentically reflects a situation in which the construct might be applied in the real-world and the actions required to produce a response are similar to those one might perform in the real-world" (Russell & Moncaleano, 2019, p. 6). When a TEI presents an authentic context but inauthentic interactions are used to produce a response, the TEI would be rated to have moderate construct fidelity. Finally, the TEI Utility Framework considers that the authenticity of the context supersedes the authenticity of the interactions, thus, whenever the context is inauthentic, regardless of the authenticity of the actions required to produce a response, the item is considered to have low construct fidelity (Russell, 2016). To inform

construct fidelity ratings, Russell and Moncaleano (2019) developed the construct fidelity coding guide shown in Appendix C which describes examples for each construct fidelity rating.

Three main factors are considered when rating the usability of a TEI: intuitiveness, layout, and functionality. Intuitiveness refers to the easiness with which a test-taker may produce a response with minimal cognitive effort (assuming the test-taker has had some training prior to exposure to the testing platform). Layout considers how the item is designed to minimize the distance between objects in the response interaction space required to produce a response. Finally, functionality refers to the extent to which the TEI is designed in a way that minimizes the number of mouse/finger selections required to produce a response (Russell, 2016). Panel members rating the usability of a TEI should consider these three factors simultaneously to produce a holistic usability rating.

A panel of three graduate students in the Measurement, Evaluation, Statistics, and Assessment (MESA) program at Boston College were trained on the use of the framework. This training was executed in four steps. First, the researcher explained the framework to the panelists, introducing the two components of construct fidelity and usability. Second, the researcher presented three TEIs as examples, leading a discussion about how to rate each component and the rationale behind the rating (high, moderate, or low). In this step, the researcher introduced the panelists to the construct fidelity coding guide (shown in Appendix C, Russell & Moncaleano, 2019). Third, the panelists worked independently to rate the construct fidelity and usability of five practice TEIs (2019 MCAS released TEIs). Finally, the researcher led a discussion among panelists to reach a consensus on the ratings and discuss any remaining questions from the panelists.

After training was completed, all panelists were asked to rate the construct fidelity and usability of the TEIs developed for this study independently. First, panelists were instructed to interact with each TEI and provide a construct fidelity rating (high, moderate, or low) for each item. Next, panelists were asked to provide an overall rating of usability (high, moderate, or low) for each TEI format (i.e., classification and rank-ordering). Panelists rated the usability of the two TEI formats as a group (rather than on an item-by-item basis) because usability ratings often depend on how test-delivery platforms implement different TEI formats more than they do on the items themselves. Once all panelists completed this process, they reconvened and discussed any discrepancies in order to reach consensus. This process led to a consensus rating of the construct fidelity of each TEI included in the instrument and the usability of both TEI formats employed.

Analytic Methods

One purpose of this dissertation is to develop a method to judge the measurement value of technology-enhanced items compared to selected-response item counterparts. To this end, all items were scored dichotomously and both classical test theory (CTT) and item response theory (IRT) approaches were employed to examine: (a) CTT difficulty, (b) CTT discrimination, (c) IRT difficulty, (d) IRT discrimination, (e) IRT item information, and (f) relative efficiency. Each of these characteristics and their calculation are described in detail. Most analyses were conducted in R using the following packages: ltm (Rizopoulos, 2018), cocor (Diedenhofen, 2016), catIrt (Nydick, 2015), and sirt (Robitzsch, 2020).

After estimating these characteristics for each item, TEIs and their MC counterparts were compared to address RQ1. Subsequently, the relationship between the psychometric characteristics of TEIs and their utility ratings was explored to answer RQ2.

Finally, to address RQ3, psychometric characteristics and utility ratings were combined to develop a protocol to judge comparative measurement value and apply it to the TEIs developed for this study.

Classical Test Theory

Under the CTT framework, two characteristics were estimated for each item, namely difficulty and discrimination. Difficulty was calculated as the proportion of participants that answered an item correctly. Discrimination was estimated as the correlation between item responses (correct/incorrect) and total test score absent of the item under consideration. Additionally, for multiple-choice items the proportion of participants that selected each option was examined. Finally, the reliability of the instrument was evaluated using Cronbach's Alpha given by

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \hat{\sigma}_i^2}{\hat{\sigma}_x^2}\right) \tag{8}$$

"where k is the number of items on the test, $\hat{\sigma}_i^2$ is the variance of item i, and $\hat{\sigma}_x^2$ is the total test variance" (Crocker & Algina, 1986, p. 138).

Item Response Theory

Two- and three-parameter logistic models are the most common psychometric models used to calibrate dichotomously-scored items in large-scale assessment programs that include TEIs. Given that guessing is not a pervasive issue for TEIs (Gifford, 2017; Huff & Sireci, 2001; Parshall & Harmes, 2014), and considering that the same model will be used to fit TEIs and MCIs in order to compare their properties, this study employed a 2-PL dichotomous IRT model to analyze participant responses. This model, provides estimates

for difficulty (b parameter), discrimination (a parameter), and information, thus allowing relative efficiency to be calculated.

The 2-PL Model. The IRT paradigm relies on logistic functions that model the probability of an item response as a function of examinees' underlying ability and item characteristics (e.g., difficulty, discrimination, guessing). In the 2-PL model the conditional probability of a dichotomous response (i.e., correct or incorrect) is modeled as a function of the difference between examinees' ability level and the difficulty of the item, weighted by the discrimination of the item. The 2-PL model takes the following statistical form:

$$p(X_{nj} = 1|\theta_n, a_j, b_j) = \frac{\exp[a_j(\theta_n - b_j)]}{1 + \exp[a_j(\theta_n - b_j)]} \tag{9}$$

where X_{nj} represents the response of examinee n to item j (correct = 1, incorrect = 0), θ_n represents the ability level of examinee n, a_j represents the item's discrimination, and b_j represents the item's difficulty. High probabilities of a correct response correspond to large positive differences between the person's ability level θ_n and the item's difficulty b_j (e.g., easy items or high-ability examinees). Conversely, when the difference between examinee ability θ_n and item difficulty b_j is negative (e.g., hard items or low-ability examinees), the probability of obtaining a correct response approaches zero. In the 2-PL model, item discrimination (a_j) acts as a weight of the difference between examinee ability and item difficulty. For highly discriminating items, small differences result in a high probability of producing a correct response. When an examinee's ability matches the difficulty of an item, the probability of obtaining a correct response is 50% (exp(0)/(1+exp(0))=0.50).

IRT models rely on assumptions of unidimensionality and local independence across items. Unidimensionality analyses were conducted using the Normal Ogive Harmonic Analysis Robust Method (NOHARM; Fraser & McDonald, 1988, 2012). Two main statistics were evaluated: the Root Mean Square (RMS) and Tanaka's Goodness of Fit Index (GFI), both of which rely on the residual matrix that results from fitting a normal ogive model to the data based on an assumed number of dimensions (one in this case). Small values of the RMS are indicative of good fit as this statistic summarizes the differences between observed and predicted covariances. In particular, the RMS statistic is often compared to a threshold equivalent to four times the reciprocal of the square root of the sample size, i.e., the "typical" standard error of the residuals (de Ayala, 2013; McDonald, 1997). Tanaka's goodness of fit index has a maximum value of 1 which indicates perfect fit, generally values above .90 are considered acceptable and values above .95 indicate good fit (McDonald, 1999). The assumption of local independence was checked by evaluating standardized LD chi-square statistics (Chen & Thissen, 1997) obtained from IRTPRO (Vector Psychometric Group, 2020). The LD statistic is concerning whenever its magnitude is larger than 10 (i.e., indicating significant dependence between items), moderate when its magnitude is between 5 and 10, and small or inconsequential when its magnitude is below 5.

Item Information. IRT models estimate the examinees' location on the ability continuum ($\hat{\theta}$). Test information functions provide a measure of these ability estimates across the ability continuum. That is, a test provides different levels of accuracy of the estimate of examinees' ability at different points on the ability continuum. As detailed in Chapter 2, the test information function corresponds to a sum of the information provided

by each item throughout the ability continuum. A general formulation for the item information function is:

$$I_j(\theta) = \frac{[p_j^{'}]^2}{p_j(1-p_j)} \tag{10}$$

where p_j is the probability of a person answering item j correctly while $p_j^{'}$ represents the first derivative of the estimated model. For the 2-PL model described in Equation 9 the first derivative corresponds to:

$$p_j^{'} = a_j p_j(1-p_j) \tag{11}$$

where a_j represents the discrimination parameter for item j. Substituting this derivative in Equation 10 produces the item information formula for the 2-PL model:

$$I_j(\theta) = a_j^2 p_j(1-p_j). \tag{12}$$

Note that item information is a function of the probability of a person answering the item correctly (p_j) and the item discrimination parameter (a_j). The 2-PL information function attains a maximum of $a^2 0.25$ at the location where a test-taker has a 50% chance of answering the item correctly (i.e., p_j=0.50). In other words, an item provides the most information at the location it discriminates the most (i.e., at b_j; de Ayala, 2013). Item information estimates at each point on the ability continuum can be plotted in a coordinate plane to produce item information curves. These curves are symmetric with respect to b_j and asymptotic (approaching zero as the magnitude of $\theta - b_j$ increases).

Relative Efficiency. The information functions provided by the estimation of the 2-PL model can be used to examine the relative efficiency of TEIs compared to multiple-choice items. Based on the literature review, there are two common ways of

estimating relative efficiency. The first method, introduced by Lord (1980) estimates a relative efficiency function as a ratio of two information functions. The second method, introduced by Jodoin (2003), calculates the ratio between the efficiency with which two items gather information per unit of time. Both of these approaches were used in this study and are described in detail next.

Lord's (1980) method for calculating relative efficiency was originally developed to compare either scores on two different tests of the same ability θ or scores resulting from scoring the same test in two different ways. Lord's method requires two steps. First, the information function for two tests is estimated using an IRT model. Second, the relative efficiency of scores from test A to scores of test B is calculated as the ratio of their information functions. However, this approach may be extended to item-level information curves:

$$RE\{A, B\} = \frac{I_A(\theta)}{I_B(\theta)} \tag{13}$$

where $I_A(\theta)$ and $I_B(\theta)$ are the information functions for items A and B respectively. Relative efficiency defined in this manner is not a point-estimate (i.e., a single value) but rather a curve. Consequently, the relative efficiency of two scores varies according to ability level. The relative efficiency curve is compared to a horizontal reference line equal to 1. Regions where $RE\{A, B\}$ is above 1, indicate item A is more efficient than item B at that range of ability level. Conversely, regions where $RE\{A, B\}$ is below 1, indicate item B is more efficient than item A at that range of ability level. Lord's relative efficiency ratio is useful for comparing information curves rather than an item statistic itself.

In contrast to Lord's (1980) method, Jodoin's (2003) approach to relative efficiency considers the time test-takers require to answer each item and yields a point-

estimate rather than a curve. Jodoin calculated the relative efficiency of one item format to another by averaging the measurement efficiency of all items that shared each format. This methodology was modified in this study to calculate relative efficiency across two items with differing response formats following four steps. First, information functions for both items in a stem-equivalent pair were estimated. Second, the information provided by each item was summarized by calculating expected information, a weighted average of information. This was accomplished by averaging the item information function (I_j) evaluated at test-takers' ability estimates ($\hat{\theta}$)[4] , i.e.,

$$E(I_j) = \frac{1}{N} \sum_{i=1}^{N} I_j(\hat{\theta}_i). \tag{14}$$

Third, the ratio of average expected information to median time spent by test-takers on the item was calculated to yield an estimate of measurement efficiency of each item (expected information per minute). Finally, relative efficiency was estimated as the ratio of TEI measurement efficiency to MCI measurement efficiency. Additionally, relative efficiency was averaged across items that shared a TEI response interaction to obtain an overall estimate of relative efficiency for that response type (i.e., classification or rank-ordering).

[4]This formula is equivalent to Equation 7 and was used as ability estimates for all test-takers had been estimated and it was assumed the sample represented the target population.

RQ1: How do the psychometric characteristics of commonly employed TEI drag-and-drop formats (classification and rank-ordering) compare to stem-equivalent multiple-choice items?

Research question 1 examined seven psychometric properties estimated for all items: (a) CTT difficulty, (b) CTT discrimination, (c) IRT difficulty, (d) IRT discrimination, (e) IRT item information, (f) expected information, and (g) measurement efficiency. Each of these properties were compared for each item pair.

CTT difficulty estimates were compared between stem-equivalent items by conducting independent samples t tests and CTT discrimination indices were compared using Fisher's (1925) test for correlations. In both occasions, a Bonferroni correction to the significance level for multiple comparisons was applied (Privitera, 2017).

Comparisons between IRT parameters were conducted by identifying item parameter drift. IRT difficulty was compared using two graphical displays. The first graphical display was an adaptation of the approach used by the TIMSS & PIRLS International Study Center to identify item parameter drift of trend items between two modes of administration or two consecutive administrations (Fishbein et al., 2020). This graphical display was constructed by plotting the difference between TEI and MCI difficulty parameters. Confidence intervals around these points were constructed according to the following formulas:

$$\begin{aligned} \text{Upper Limit} &= b_{TEI} - b_{MCI} + SE(b_{TEI} - b_{MCI}) \cdot Z_b \\ \text{Lower Limit} &= b_{TEI} - b_{MCI} - SE(b_{TEI} - b_{MCI}) \cdot Z_b \end{aligned} \tag{15}$$

where b_{TEI} and b_{MCI} are the difficulty parameters for the TEI and the stem-equivalent MCI version respectively, $SE(b_{TEI} - b_{MCI})$ is the standard error of the differences between difficulty parameters of TEIs and MCIs, and Z_b is the 95% critical value of the z distribution corrected for multiple comparisons. Using this approach differences between difficulty parameters larger than 2 logits are considered concerning.

The second graphical display used to compare difficulty parameters followed the "3-sigma IRT" approach (Gaertner & Briggs, 2009). The TEI and MCI difficulty parameters were plotted against each other and the standard deviation (SD) line was used as the line of best fit. The SD line given by the following equation:

$$y = \frac{sd(b_{TEI})}{sd(b_{MCI})}x \tag{16}$$

where $sd(b_{TEI})$ and $sd(b_{MCI})$ are the standard deviations of the difficulty parameters of TEIs and MCIs respectively. A confidence interval was also constructed around the SD line at a distance of 3 times the standard deviation of the perpendicular distances between each point and the SD line. Under this approach, items beyond this confidence region are interpreted as having concerning item parameter drift. On this graphical display the $y = x$ line was also drawn as a reference to compare difficulty parameters at face value (i.e., larger than, less than).

IRT discrimination parameters could not be compared using the first graphical display given that discrimination parameters are not interpreted in a logits metric. However, discrimination parameters were compared using the second graphical display described above. The SD line and the confidence region were constructed following the

same calculations shown in Equations 15 and 16 using discrimination parameters (i.e., a_{TEI} and a_{MCI}).

Item information curves for TEIs were compared to item information curves of their multiple-choice counterparts using two approaches: (a) plotting them on the same coordinate plane and inspecting visually the extent to which they overlap and (b) applying Lord's (1980) relative efficiency ratio method. Given the asymptotic nature of item information curves, it is rare that an item provides more information than another across the full ability continuum. In other words, item information functions often intersect, indicating that while one item might provide more information than another in a specific range, the opposite is true in a different interval. To characterize the comparisons made in this study the ability continuum was divided in three easily interpretable intervals informed by Qian et al.'s (2017) study: (a) $\theta < -1$ or low-ability examinees, (b) $-1 \leq \theta \leq 1$ or average-ability examinees, and (c) $1 < \theta$ or high-ability examinees. For each item pair and for each of these intervals the amount of information provided by TEIs and MCIs was compared visually. Additionally, Lord's relative efficiency ratio function was plotted in a coordinate plane and interpreted using the three ability ranges described above for information curves. For each of these intervals, the item in a pair that appeared to be more efficient was noted.

Expected information and measurement efficiency estimates were directly comparable because they are point-estimates. Following Jodoin's (2003) methodology, ratios of expected information and measurement efficiency were calculated (referred to henceforth as relative expected information and relative measurement efficiency respectively). Values of these ratios larger than 1 indicated that the TEI version

performed better than its MCI counterpart and values below 1 indicated the MCI performed better. Additionally, in his analyses, Jodoin considered relative efficiency values larger than 2 to be meaningful. This criterion was applied to both ratios calculated in this study to identify meaningful differences in expected information and measurement efficiency between stem-equivalent items.

RQ2: What is the relationship between the utility of TEI drag-and-drop formats (classification and rank-ordering) and their psychometric item characteristics?

Research question 2 examined the relationship between the utility ratings of the TEIs used in the instrument as determined by a panel of specialists and the psychometric characteristics calculated for each item to address research question 1. To explore this relationship, all TEIs were first classified according to their construct fidelity rating (high fidelity, moderate fidelity, or no fidelity). Within each of these groups, the psychometric properties of the items were analyzed in a holistic manner to identify any patterns. Differences in patterns of the results of the comparisons of the psychometric properties between construct fidelity rating categories were recorded. The identified patterns were used to characterize each of the construct fidelity rating levels. Moreover, acknowledging that the features of the platform chosen for the delivery of the instrument (Qualtrics) were out of control from the researcher, the relationships between the ratings of usability and item psychometric properties were analyzed independently.

RQ3: How can TEI psychometric properties and utility ratings be combined to develop a standardized protocol to judge the comparative measurement value of TEIs relative to stem-equivalent MC items?

This dissertation aims to provide a standardized and replicable methodology to evaluate whether a TEI provides "better" measurement of a construct compared to a stem-equivalent multiple-choice counterpart. The intent of this judgment is to inform decisions about the use of a TEI instead of a selected-response item format. To this end, this dissertation estimated several psychometric properties of TEIs and their MC counterpart (i.e., stem-equivalent) and proposes a protocol that examines these properties to evaluate the comparative measurement value of a TEI. Comparative measurement value is an overall judgment regarding the benefit of using one item format versus an alternate format with respect to increasing construct fidelity and improving psychometric characteristics. As detailed in chapter 2, multiple psychometric properties have been used to evaluate the quality of traditional item formats based on standards and thresholds commonly accepted in the field of educational measurement. The protocol proposed in this dissertation extends the use of psychometric properties to judge the comparative measurement value of technology-enhanced items.

As described in this chapter, nine item characteristics were estimated for each item as part of research questions 1 and 2 (seven for RQ1 and two for RQ2). Research question 3 evaluated these characteristics to identify the best indicators to judge the comparative measurement value (CMV) of TEIs relative to stem-equivalent multiple-choice items. Once these indicators were chosen, they were examined to identify desirable and undesirable outcomes and inform the order in which these indicators should be evaluated

within the protocol. These indicators were then combined into a step-wise hierarchical decision protocol that provides a judgment about the impact on comparative measurement value judgment resulting from the use of the TEI format.

From a practical standpoint, the comparative measurement value protocol aims to provide an answer to the following question: is there value in using a given TEI format instead of a traditional MC version? Consequently, increased CMV is an indication that the TEI format was superior according to at least one indicator in the protocol and thus it is the recommended format. In contrast, decreased CMV indicates that an egregious undesirable outcome was observed, thus recommending the MC format. Finally, if no impact on CMV is evident the protocol recommends a multiple-choice format given that no clear benefits of using a TEI format are observed. In this sense, the protocol provides a clear representation of the tradeoffs between choosing a TEI or an MCI.

The resulting protocol was applied separately to each technology-enhanced item developed in this study. First, each indicator was examined independently comparing the characteristics of the TEI version to those of the MCI version. The results of these comparisons were then pooled using the protocol to arrive at an overall judgment of comparative measurement value. Moreover, once comparative measurement value judgments were obtained for all items, TEIs that shared a common response interaction (classification or rank-ordering) were examined to evaluate whether there was an association between comparative measurement value and item response format.

Chapter 4 - Results

This chapter presents results of the procedures described in Chapter 3. The chapter begins by describing changes made to the data collection instrument based on two pilot administrations. Subsequently, results of the operational administration are presented, including characteristics of the sample and analyses of omitted responses and timing data. Then this chapter presents comparisons across stem-equivalent TEIs and MCIs based on classical test theory and item response theory item parameters, item information curves, and efficiency. Results of TEI utility ratings are also presented in this chapter. This chapter concludes with a summary of the results and how they inform the three main research questions of this study.

Instrument Development

The primary purpose of this study is to develop a protocol to evaluate the comparative measurement value of technology-enhanced items and stem-equivalent multiple-choice items with the intent of informing decisions about the use of the former instead of the latter. To accomplish this, two forms of an 18-item data collection instrument were developed. The first six items of the instrument comprised a common block presented to all participants in this study, the remaining 12 items, organized in two blocks, differed on their response interactions across forms. Therefore, each instrument form included the common block, a drag-and-drop block, and a multiple-choice block. Form A included blocks TEI-1 and MCI-2 while Form B included blocks TEI-2 and MCI-1. Henceforth, item blocks TEI-1 and MCI-1 are referred to as item set 1, as they contain stem-equivalent items, and blocks TEI-2 and MCI-2 constitute item set 2.

The use of stem-equivalent pairs of technology-enhanced and multiple-choice items is the cornerstone of this study, as it prevents comparisons from being confounded with differences in content across item formats (a common limitation of item format comparison studies). The items in the data collection instrument were intended to have moderate difficulties. Because the sample contained adults recruited through MTurk, it was anticipated that extremely difficult items would demotivate participants while extremely easy items would produce limited score variability. Two pilots were conducted to evaluate item difficulty (percent correct) and flag items for review or replacement. Results of both pilots are described in the following sections.

First Pilot

The first pilot began by developing 30 items. Six items were developed to build the common block: two multiple-choice items, two classification drag-and-drop items, and two drag-and-drop rank-ordering items. Twelve drag-and-drop TEIs (six classification items and six rank-ordering items) and 12 multiple-choice items that were stem-equivalent to each of the TEIs were developed to populate the remaining blocks. All items were subjected to a content review. This initial set of 30 items assessed concepts covered in high school statistics including: measures of central tendency, properties of frequency distributions, and interpretation of graphical data displays.

A total of 121 participants were recruited through MTurk for the pilot. Upon inspection of the attention control items, 62 responses were discarded for incorrect answers, resulting in a 51% attrition rate due to inattentiveness. Based on the remaining 59 valid responses, the percent of correct responses for each item was calculated. Results are shown in Table 4.1 organized by item block and instrument form (items are shown in

the order within a block in which they were presented to participants). To scaffold the decision process, items were flagged according to their difficulty level (below 16%, between 16% and 30%, between 70% and 83%, and above 83%).

Table 4.1

First pilot results

Item	Difficulty	Item	Difficulty	Decision
		Common Block		
MC1	40.7			Keep
CL1	66.1			Keep
RO1	75.0†			Keep
MC2	40.7			Keep
CL2	13.6**			Replace
RO2	20.3*			Replace
		Item Set 1		
TEI-1		MCI-1		
CL3	0**	MCCL3	50.0	Replace
RO3	0**	MCRO3	26.7*	Replace
CL4	31.0	MCCL4	67.0	Keep
RO4	31.0	MCR04	16.7*	Keep
CL5	44.8	MCCL5	40.0	Keep
RO5	24.1*	MCRO5	30.0*	Replace
		Item Set 2		
TEI-2		MCI-2		
CL6	40.0	MCCL6	41.4	Keep
CL7	10.0**	MCCL7	31.0	Replace
RO6	0**	MCRO6	55.2	Replace
CL8	30.0	MCCL8	41.4	Keep
RO7	86.7††	MCRO7	68.8	Keep
RO8	20.0*	MCRO8	13.8**	Replace

Note. **Below 16%. *Between 16% and 30%. †Between 70% and 83%. ††Above 83%.

Six items were flagged for replacement due to their extreme difficulty (less than 16% correct responses): CL2, CL3, RO3, CL7, RO6, and MCRO8. One item was flagged for being extremely easy (more than 83% correct responses): RO7. Seven items were flagged for revision due to their minimally acceptable difficulty (between 16% and 30% or

70% and 83% correct responses): RO1, RO2, MCRO3, MCRO4, RO5, MCRO5, and RO8. Based on these results, the following decisions were made:

1. The instrument as a whole was deemed too difficult for the target population, consequently items that were flagged for being too easy were retained (RO1 and RO7).

2. Both flagged items in the common block were replaced (CL2 and RO2). Although, item RO2 was acceptable, an item with a higher percent correct was desirable given the use of the common block to develop a common scale.

3. All item pairs that included an item flagged for extreme difficulty (**) were replaced. Item pairs where one item was flagged for review (*) and the other was not, were retained. Finally, item pairs where both items were flagged for review were replaced.

These decisions resulted in 14 items selected for replacement, two single items from the common block and six item pairs.

Two main conclusions were drawn from the results of this pilot: (a) attrition due to careless responses was high and (b) the data collection instrument was too difficult for the target population. It was hypothesized that these two conclusions might be related, the high difficulty of the instrument could explain a lack of motivation and thus careless responses. For this reason, the content of the data collection instrument was broadened by replacing some flagged items with adaptations of 5th-, 6th-, 7th-, and 8th-grade mathematics and science MCAS 2019 released items (MA-DESE, n.d.-a, n.d.-b). Additionally, item CL8 and its counterpart MCCL8 were replaced. This item pair

presented test-takers with a set of three box plots and asked them to classify them according to the skewness of the data they portrayed. Despite their acceptable percent correct responses, these items were removed following the recommendation of a content expert who suggested these items seemed to assess recall of vocabulary rather than statistical understanding.

Second Pilot

Given the extensive revisions to the data collection instrument, a second pilot was conducted. At this stage, the data collection instrument comprised 11 statistics items, five science items, and two mathematics items. The statistics items covered all three sub-domains included in the prior version: measures of central tendency, properties of frequency distributions, and interpretation of graphical data displays. The science items covered the following subdomains (as defined by the Massachusetts standards the adapted items addressed): technological systems, earth's systems, matter and its interactions, and forces and interactions. The two adapted mathematics items addressed number properties and geometry.

In addition to the changes to the instrument, the MTurk recruitment message, the informed consent, and the instructions were edited to advise participants their task might not be approved (and thus they might not be compensated) if their responses showed evidence of carelessness. For this second pilot 157 participants were recruited. Sixty-five responses were rejected due to carelessness (35 due to incorrect responses to attention control items and completing the instrument in less than 5 minutes, 22 due to incorrect responses to attention control items despite an acceptable instrument completion time, and 18 due to incoherent open-ended responses) resulting in 92 valid responses.

Consequently, this pilot showed a 41% attrition rate, marking an improvement from the first pilot. The percent correct responses per item was once again calculated. Results are shown in Table 4.2 organized by item block and instrument form (items are shown in the order within a block in which they were presented to participants).

Table 4.2

Second pilot results

Item	Difficulty	Item	Difficulty	Decision
		Common Block		
MC1	32.6			Keep
CL1	51.1			Keep
RO1	55.4			Keep
MC2	27.2*			Keep
CL2‡	44.6			Keep
RO2‡	10.7**			Review
		Item Set 1		
TEI-1		MCI-1		
CL3‡	60.4	MCCL3‡	52.3	Keep
RO3‡	35.4	MCRO3‡	34.1	Keep
CL4	20.8*	MCCL4	56.8	Keep
RO4	29.2	MCRO4	22.7*	Keep
CL5	27.1*	MCCL5	45.5	Keep
RO5‡	14.6**	MCRO5‡	18.2*	Review
		Item Set 2		
TEI-2		MCI-2		
CL6	29.6*	MCCL6	31.9	Keep
CL7	40.9	MCCL7	47.9	Keep
RO6	65.9	MCRO6	54.2	Keep
RO7‡	9.1**	MCRO7‡	29.2*	Review
CL8‡	72.7†	MCCL8‡	56.3	Keep
RO8‡	59.1	MCRO8‡	50.0	Keep

Note. ‡New item. **Below 16%. *Between 16% and 30%. †Between 70% and 83%. ††Above 83%.

Results of the second pilot flagged three items for extreme difficulties (RO2, RO5, and RO7) and seven items for minimally acceptable difficulties. No items were replaced at

this stage, but rather reviewed and edited as necessary. The following decisions were made:

1. Items CL4, MCRO4, CL5, and CL6 were left intact.

2. Item RO2 was found to have a confusing prompt. The prompt was edited.

3. Items RO5 and MCRO5 required participants to calculate the density of three objects based on provided mass and volume values. As this process required remembering a formula it was revised to include a scaffold.

4. Item RO7 was found to be confusing because it shared the same graphical stimulus as the preceding item (RO6) and both questions were worded similarly. Analysis of the data showed participants often repeated their answer to RO6/MCRO6 in item RO7/MCRO7. The prompts were modified and these items were separated into two different blocks despite sharing the same stimulus.

The final data collection instrument comprised 11 statistics, five science, and two mathematics items addressing the subdomains described earlier. Before engaging with the data collection instrument, participants were asked to answer a brief background questionnaire that inquired about their level of education. Additionally, after completing the instrument, participants were shown a closing survey which included an open-ended question asking participants about their previous experience with statistics (e.g., coursework) and a question about whether they took breaks during the task. The background questionnaire, the final instrument, and the closing survey are presented in Appendix D.

Operational Administration and Sample Characteristics

After finalizing the data collection instrument based on findings from the two pilot studies, the instrument was administered to 900 participants recruited through MTurk, with the goal of obtaining at least 300 valid responses per item. Of the 901 responses received 307 were removed for carelessness indicated by incorrect attention control items, short response times, incoherent answers to open-ended response items, or answers that appeared automatic. A total of 594 responses were deemed valid for further analyses, corresponding to an attrition rate of 34%.

As described in Chapter 3, participants were randomly assigned to one of two possible forms (A —Blocks TEI-1 and MCI-2 or B —Blocks TEI-2 and MCI-1) and one of two possible block orders (1 —Common block first, TEI block second, MCI block third, or 2 —Common block first, MCI block second, TEI block third). Table 4.3 shows the number of participants who answered each form.

Table 4.3

Number of participants who answered each form of data collection instrument

Order	Form	
	A (TEI-1 and MCI-2)	B (TEI-2 and MCI-1)
1 (CB - TEI - MCI)	137	146
2 (CB - MCI - TEI)	146	165
Total	283	311

Table 4.4 displays a description of the sample based on responses to the background questionnaire which participants answered prior to engaging with the data

collection instrument. All participants indicated they understood English very well or well. A small percentage of participants reported not having a higher-education degree (16%) and participants who indicated having a higher-education degree were evenly split between Mathematics or Statistics degrees and other degrees. The majority of participants (54%) have enrolled in at least one higher-education statistics course in the past, and a small percent (16%) reported having experience teaching statistics, mathematics, or science at any academic level.

Table 4.4

Sample characteristics

Variable	Percent
Understanding of English	
Very well	97
Well	3
Not very well	0
Higher-Education Degree	
Mathematics or Statistics	40
Other	44
No degree	16
Ever enrolled in higher-education statistics courses	
Yes	54
No	46
Experience teaching statistics, mathematics or science at any academic level	
Yes	16
No	84

Omitted Responses

Participants were allowed to skip items within the data collection instrument. A total of 106 participants (17.85%) skipped (or did not respond to) at least one item. Table 4.5 shows the distribution of participants who omitted items by the number of items omitted

and instrument form. Results indicate that the majority of participants who skipped items only skipped a single item. Moreover, the percent of people who skipped at least one item was similar across forms (17% for Form A and 18% for Form B).

Table 4.5

Number of participants who omitted responses per form

Items Omitted	Form A	Form B	Total
1	42	41	83
2	3	12	15
3	1	1	2
4		1	1
5	2	3	5
Total	48	58	106

Table 4.6 shows the prevalence of omitted responses for each item and form of the instrument. For each item, the proportion of omitted responses is reported based on the total number of observations available for each item (i.e., the number of participants who answered each form). In total, 61 omitted responses were observed in Form A and 87 occurred in Form B, corresponding to less than 2% of the data gathered per form. The last item in the common block (RO2) was the item that showed the highest number of omitted responses (56).

The pattern of missing responses suggests that participants were more likely to skip technology-enhanced items than multiple-choice items. All TEIs showed at least one case of an omitted response (1.46% missing responses on average) while only three MCIs showed cases of omitted responses (0.05% missing responses on average). This may be associated with the additional effort involved in interacting with a TEI compared to a traditional multiple-choice item. To examine the extent to which missing data occurred

randomly, Little's MCAR test was performed in SPSS (Enders, 2010; Little, 1988). Chi-square statistics were non-significant for both forms indicating that missing data occurred at random—$\chi^2_A(178, N = 283) = 164.88$, $p = .751$ and $\chi^2_B(241, N = 311) = 242.68$, $p = .457$. Therefore, despite the apparent relationship between omitted responses and item type, omitted responses were considered random and scored as incorrect responses.

Table 4.6

Omitted responses for each item per form

Form A (N=283)			Form B (N=311)		
Item	Omitted	% Omitted	Item	Omitted	% Omitted
Common			Common		
MC1			MC1	1	0.32
MC2			MC2		
CL1	3	1.06	CL1	6	1.93
CL2	3	1.06	CL2	4	1.29
RO1	1	0.35	RO1	2	0.64
RO2	30	10.60	RO2	26	8.36
TEI-1			MCI-1		
CL3	2	0.71	MCCL3		
CL4	2	0.71	MCCL4	1	0.32
CL5	2	0.71	MCCL5		
RO3	8	2.83	MCRO3		
RO4	5	1.77	MCR04		
RO5	4	1.41	MCRO5		
MCI-2			TEI-2		
MCCL6			CL6	4	1.29
MCCL7			CL7	3	0.96
MCCL8			CL8	3	0.96
MCRO6	1	0.35	RO6	17	5.47
MCRO7			RO7	14	4.50
MCRO8			RO8	6	1.93
Total	61	1.19[a]	Total	87	1.55[b]

[a]Form A total observations = 5094. [b]Form B total observations = 5598.

Timing Statistics

This section discusses (a) the total time taken to complete the full instrument, (b) the total time taken to complete each block, and (c) the time taken to complete each item. To characterize the distribution of timing data at each of these levels the mean, standard deviation, minimum, and maximum times are reported. Considering test-taking timing data is often positively skewed due to the presence of outliers (i.e., participants who invested too much time; van der Linden, 2006; Weeks et al., 2016), the median time is also reported as the preferred measure of central tendency. Results of statistical significance tests comparing time spent on different item formats are presented following the discussion of these descriptive statistics at each level. This section concludes with a discussion of the impact of omitted responses on timing statistics.

Instrument-level Timing Statistics

Table 4.7 presents timing statistics at the instrument level. Results are also discriminated by the form participants answered (A or B) as well as the block order participants experienced (1 or 2). Results suggest that participants spent, on average, about 20 minutes answering all items in the instrument (not including the background questionnaire or the closing survey). Results also indicate that participants who answered Form A spent marginally more time than those answering Form B (about a minute longer). The difference in time spent in Form A ($M = 21.56, SD = 10.86$) and Form B ($M = 20.63, SD = 10.17$) was not significant, $t(592) = 1.08, p = .281$. Participants who were assigned Form A2 took, on average, about 2 minutes longer to complete the instrument. However, a one-way analysis of variance (ANOVA) conducted to determine whether this difference was significant indicated that there were no significant differences

across the mean times spent by participants in each of the four possible forms and order of blocks, $F(3, 590) = 0.71$, $p = .546$.

Table 4.7

Overall instrument timing descriptive statistics

Variable	N	M	SD	Min	Max	Mdn
All participants	594	21.08	10.51	2.34	79.13	19.50
By instrument form						
A	283	21.56	10.86	2.43	79.13	20.59
B	311	20.63	10.17	2.34	56.58	19.12
By form and block order						
A1	137	20.94	11.20	2.43	79.13	19.29
A2	146	22.15	10.54	2.74	48.40	21.13
B1	146	20.74	10.56	2.34	56.58	19.11
B2	165	20.54	9.85	3.50	48.34	19.10

Note. All times are shown in minutes.
$^*p < .05$.

A closer look at the distribution of the total time spent by participants revealed that 22.2% finished in less than 15 minutes, 51.2% spent between 15 and 30 minutes, 20.2% spent between 30 and 45 minutes, 6.1% spent between 45 minutes and 1 hour, and 0.3% spent more than 1 hour. Overall, about 73% of participants spent less than 30 minutes which corresponds with the expected time required to complete the instrument.

Finally, the closing survey included a question asking participants whether they took any breaks while working on the instrument. Results showed that 557 participants did not take a break (93.8%) while 37 did take a break (6.2%). On average, participants who did not take breaks took approximately a minute longer to answer the instrument compared to participants who took breaks. This counter-intuitive difference may be explained due to the disproportionate difference in sample sizes and by the fact that the

only instances of participants taking more than one hour corresponded to people who did not take breaks. However, the difference of time spent between participants who took breaks ($M = 20.15, SD = 11.65$) and participants who did not ($M = 21.13, SD = 10.44$) was not found to be statistically significant, $t(592) = 0.55, p = .585$.

Block-level Timing Statistics

Descriptive statistics for the total amount of time participants spent in each block are shown in Table 4.8. Median times ranged from 5.27 to 6.36 minutes indicating that all item blocks required approximately 6 minutes to complete. Both the means and the medians suggest that participants spent marginally more time on the blocks corresponding to item set 1 (TEI-1 and MCI-1) than on blocks of item set 2. Results of a one-way ANOVA indicated that differences in mean time spent across blocks was statistically significant, $F(4, 1777) = 2.89, p = .021$, however Scheffe's post-hoc tests[5] did not reveal any statistically significant pair-wise differences.

Table 4.8

Block-level timing descriptive statistics

Block	*N*	*M*	*SD*	Min	Max	Mdn
Common	594	6.95	4.12	0.73	34.77	6.09
TEI-1	283	6.93	4.70	0.62	43.77	6.15
MCI-1	311	7.00	4.54	0.35	24.25	6.36
TEI-2	283	6.08	3.80	0.44	22.99	5.27
MCI-2	311	6.49	4.38	0.25	29.58	5.81

Note. All times are shown in minutes.

[5]Scheffe's test is considered more conservative than other post-hoc tests such as Tukey's Honestly Significant Difference and Bonferroni's procedure and more appropriate for groups with unequal sample sizes (Privitera, 2017).

Item-level Timing Statistics

Descriptive statistics for item-level timing data are presented in Table 4.9. Results show that, on average, participants spent less than 2 minutes per item. Time spent on an item ranged between 0 minutes (participant skipped the item without reading it) and 28.15 minutes. Although the minimum time spent per item is fairly consistent across items, the maximum time spent varies considerably: four items showed a maximum time of 20 minutes or more, while 12 items showed a maximum time between 10 and 15 minutes. Although not included in Table 4.9, median response times for attention control items were 24 seconds for AC1 and 16 seconds for AC2, confirming that these items had a low cognitive bearing and did not hinder participants on their progress through the instrument.

Twelve independent means t tests were conducted comparing the average time spent by participants on each TEI and the corresponding multiple-choice counter parts (e.g., CL3 vs. MCCL3). A Bonferroni correction for multiple comparisons was used to set the significance threshold ($\alpha = 0.05/12 = 0.004$) and achieve an experiment-wise alpha level of 0.05. Results revealed only one statistically significant difference between items CL7 ($M = 0.57, SD = 0.41$) and MCCL7 ($M = 0.92, SD = 1.42$), $t(592) = -4.05, p < .001$. Participants answering item MCCL7 took on average 35 seconds longer than participants answering item CL7. In addition to being statistically significant, this difference is also meaningful as it constitutes over 50% of the mean and median time participants spent on item CL7.

Table 4.9

Item-level timing descriptive statistics

Item	*M*	*SD*	Min	Max	Mdn
		Common Block			
MC1	1.72	2.11	0.05	28.15	1.29
MC2	0.91	1.15	0.04	14.68	0.66
CL1	1.68	1.27	0.13	11.22	1.41
CL2	0.89	0.93	0.04	18.07	0.72
RO1	0.79	0.60	0.04	9.23	0.66
RO2	0.97	0.73	0.05	6.57	0.81
		Item Set 1			
TEI-1					
CL3	1.04	0.97	0.02	9.31	0.81
CL4	1.72	1.32	0.04	9.18	1.51
CL5	1.48	1.27	0.05	10.40	1.24
RO3	0.96	1.09	0.04	11.15	0.68
RO4	1.06	1.14	0.05	6.90	0.66
RO5	0.67	1.21	0.04	15.96	0.39
MCI-1					
MCCL3	1.14	1.28	0.06	16.59	0.87
MCCL4	1.62	1.27	0.04	7.77	1.39
MCCL5	1.45	1.42	0.03	14.40	1.20
MCRO3	1.05	1.03	0.03	7.96	0.74
MCRO4	1.14	1.32	0.03	14.47	0.77
MCRO5	0.60	0.70	0.03	9.08	0.45
		Item Set 2			
TEI-2					
CL6	1.50	1.32	0.06	12.33	1.26
CL7	0.57	0.41	0.06	4.80	0.48
CL8	0.86	0.59	0.06	5.14	0.73
RO6	1.14	1.25	0.05	10.84	0.81
RO7	0.86	1.70	0.04	20.00	0.52
RO8	1.14	1.10	0.03	8.94	0.91
MCI-2					
MCCL6	1.70	2.11	0.02	24.05	1.33
MCCL7	0.92	1.42	0.00	18.68	0.68
MCCL8	0.88	0.66	0.01	4.18	0.72
MCRO6	1.25	1.88	0.03	23.73	0.87
MCRO7	0.63	0.58	0.04	4.28	0.46
MCRO8	1.12	1.07	0.02	6.80	0.82

Note. All times are shown in minutes.

Item-level timing results do not show clear patterns regarding whether participants spent more time working on a specific item format. Three independent mean t tests were conducted on item-level mean times comparing (a) technology-enhanced items and stem-equivalent multiple-choice items (TEIs vs. MCIs), (b) classification drag-and-drop items and stem-equivalent multiple-choice items (CLs vs. MCCLs), and (c) rank-ordering drag-and-drop items and stem-equivalent multiple-choice items (ROs vs. MCROs). Although several observations (item-level mean times) were obtained from the same sample of people (e.g., the same participants answered items CL3, CL4, and CL5) these mean times were assumed independent as the instrument was not timed. In other words, as participants could take as long as they wished in each item, there is no association between the time spent on two different items. The results of these three t tests are shown in Table 4.10. For all three tests, a correction for multiple comparisons was used to set the significance threshold ($\alpha = 0.05/3 = 0.016$). Overall, results indicated there were no significant differences in time spent across item response formats.

Table 4.10

Results of independent means t tests on item-level mean times by response format

Comparison	N	M	MD	t	df	p	95% CI
TEI	12	1.08					
MCI	12	1.13	-0.04	-0.29	22	0.777	[-0.33, 0.25]
CL	6	1.20					
MCCL	6	1.28	-0.09	-0.39	10	0.708	[-0.61, 0.43]
RO	6	0.97					
MCRO	6	0.97	0.01	0.06	10	0.953	[-0.30, 0.31]

Note. MD = Mean Difference. Common block items were not included in these t tests. $p < .016$.

Omitted Responses and Timing Statistics

All timing data results discussed so far included timing data corresponding to responses that were omitted. It would be expected that participants who skipped items spent a short amount of time reading the prompt before choosing not to respond and move on. On average, participants who skipped an item spent 44 seconds on the screen corresponding to the omitted response. Times associated with omitted responses ranged approximately from 1.5 seconds to 2.5 minutes. The timing records associated with omitted responses may lower average times spent per item. In particular, time estimates for TEIs may be biased as omitted responses appeared to be more common among technology-enhanced items than multiple-choice items.

Timing statistics were recalculated by removing times associated with omitted responses and all analyses discussed in this section so far were repeated. All conclusions presented above held equally for timing data when times associated with omitted responses were removed. This is not surprising given the low prevalence of omitted responses (less than 2% in each instrument form).

Comparison of Item Characteristics

This section presents estimates of the item characteristics of interest identified in Chapter 3. This section begins with a discussion of classical test theory statistics (difficulty, discrimination, and multiple-choice response patterns) followed by the parameter estimates of a two-parameter item response theory model (difficulty and discrimination). Based on the latter, information analyses are conducted and item efficiency is examined.

Throughout this section stem-equivalent technology-enhanced and multiple-choice items are compared.

Classical Test Theory

This section presents descriptive statistics for each item including, difficulty, and discrimination. Additionally, for multiple-choice items, response patterns are examined. Items are organized by block and item response format. As described above, participants were allowed to skip items and omitted responses were scored incorrect. This section concludes with a discussion of the reliability and unidimensionality of the instrument.

Common Block. Table 4.11 shows descriptive statistics of common block items, indicating their corresponding content area, difficulty (percent correct responses), discrimination (correlation with total score absent the item), and, for multiple-choice items, the percent of respondents who selected each of the six available options. Results indicate that all items showed acceptable percent correct responses (between 16% and 83%). All items in the common block displayed acceptable discrimination statistics ranging between 0.40 and 0.55. The distribution of responses for each option in the two multiple-choice items revealed that the correct option was chosen by the majority of the participants. Moreover, for both items, one distractor was highly attractive (option D in both) while two other distractors were selected by less than 5% of the respondents (options A and B for item MC1 and options E and F for item MC2).

Table 4.11

Item descriptive statistics for the Common Block

Item	Content	Diff	Disc	A	B	C	D	E	F
MC1	Statistics	46.5	.46	3.4	4.6	5.1	26.6	**46.5**	13.8
MC2	Statistics	40.6	.43	7.6	**40.6**	12.3	36.0	2.9	0.7
CL1	Statistics	68.5	.50						
CL2	Science	64.1	.46						
RO1	Statistics	70.5	.55						
RO2	Science	29.5	.40						

Note. For MC items, the correct answer is underlined and the most common response is in bold.

Item Set 1. Descriptive statistics for item set 1 (Table 4.12) indicate that items RO3 and RO4 were the most difficult items in block TEI-1. This patterns also holds for their stem-equivalent counterparts in block MCI-1. There is no clear pattern regarding a specific TEI response format being more or less difficult than the corresponding multiple-choice pairs in this item set. The statistical significance of differences in difficulty between stem-equivalent items was evaluated by conducting independent-samples t tests (significance threshold adjusted for 12 multiple comparisons —six tests for each item set: $\alpha = 0.004$). Results indicated significant differences in difficulty between items in two pairs, CL3-MCCL3 and CL4-MCCL4 ($p < .004$). Discrimination values revealed four items with undesirable discrimination: RO3 and RO4 and their multiple-choice counterparts MCRO3 and MCRO4. Fisher's (1925) test for statistical significance between correlations of independent samples was applied to discrimination indices of stem-equivalent items (cocor package; Diedenhofen, 2016). No statistically significant results were found. Table 4.12 also shows the percent of respondents that chose each option for multiple-choice items in block MCI-1. For four of the six items the correct response was

chosen by the majority of the participants. However, for items MCRO3 and MCRO4 an incorrect option was selected by the highest percentage of participants (Option C in both).

Table 4.12

Item descriptive statistics for Item Set 1 (Blocks TEI-1 and MCI-1)

Item	Content	Diff[a]	Disc	A	B	C	D	E	F
				TEI-1					
CL3	Science	68.2^{-}	.28^{-}						
CL4	Statistics	38.2^{+}	.44^{+}						
CL5	Statistics	38.9^{+}	.49^{-}						
RO3	Statistics	27.6^{+}	-.17^{-}						
RO4	Statistics	25.1^{-}	$< .001^{+}$						
RO5	Science	66.1^{-}	.30^{-}						
				MCI-1					
MCCL3	Science	56.6*	.35	**56.6**	5.5	6.4	27.7	3.2	0.6
MCCL4	Statistics	70.4*	.36	13.9	3.6	8.4	**70.7**[b]	2.3	1.3
MCCL5	Statistics	48.2	.52	10.0	**48.2**	9.7	14.5	10.3	7.4
MCRO3	Statistics	29.3	-.05	6.8	29.3	**41.8**	13.5	5.8	2.9
MCRO4	Statistics	21.5	-.07	15.1	20.6	**24.8**	12.5	21.5	5.5
MCRO5	Science	55.3	.32	9.0	**55.3**	19.6	11.3	1.9	2.9

Note. For MC items, the correct answer is underlined and the most common response is in bold.

[a] The magnitude of the difficulty estimate and its interpretation have an inverse relationship. A smaller difficulty estimate indicates the TEI was more difficult than the MCI ($^{+}$) and a larger difficulty estimate indicates the TEI was easier than the MCI ($^{-}$).

[b] Percent correct responses and percent of participants choosing the correct response differ due to omitted responses.

$^{+}$TEI statistic was higher than MCI. $^{-}$TEI statistic was lower than MCI.

* $p < .004$.

Rank-ordering items RO3 and RO4, and their multiple-choice counterparts (MCRO3 and MCRO4) showed undesirable item characteristics: extremely low or negative discrimination and high difficulty. Moreover, for items MCRO3 and MCRO4 an incorrect option was chosen more often than the correct response. A review of the wording of the prompts and response options to these items did not reveal any issues that would

explain these response patterns. However, a closer look at the responses provided by participants to both of these item pairs revealed significant agreement across item formats. For example, as shown in Table 4.12, the most common response to item MCRO3 was option C and analyses showed that the order of objects conveyed in this option was the most common order participants who answered the TEI version provided as a response. In addition, the percentage of participants who chose the most common (incorrect) order as their response were fairly similar across formats. Continuing the example, about 42% participants chose option C in item MCRO3 while about 37% of participants answering item RO3 provided the order of objects conveyed in this option as a response. Overall, these patterns suggest that the undesirable item characteristics observed were the product of misconceptions rather than issues with the item prompts.

Item Set 2. Descriptive statistics for item set 2 (Table 4.13) indicate that all six technology-enhanced items had higher percentages of correct responses than their multiple-choice counterparts. The hardest items in these blocks were consistent across item response formats (CL6-MCCL6 and RO6-MCRO6). Independent means t tests conducted across stem-equivalent items revealed statistically significant differences in difficulties for two item pairs, CL8-MCCL8 and RO6-MCRO6 ($p < .004$), in both cases indicating that the TEI version was significantly easier than the multiple-choice version. Fisher's (1925) independent samples tests for correlation coefficients were conducted comparing discrimination statistics for each item pair, however, no statistically significant differences were found. For all multiple-choice items, the correct option was chosen by the majority of participants. For items MCCL7, MCRO6, and MCRO7, option B was the most common incorrect response.

Table 4.13

Item descriptive statistics for Item Set 2 (Blocks TEI-2 and MCI-2)

Item	Content	Diff[a]	Disc	A	B	C	D	E	F
				TEI-2					
CL6	Statistics	40.5$^-$	.54$^+$						
CL7	Science	58.8$^-$	.53$^+$						
CL8	Statistics	74.9$^-$	.37$^-$						
RO6	Statistics	39.6$^-$	.38$^-$						
RO7	Math	52.4$^-$	.44$^+$						
RO8	Math	73.6$^-$	.36$^-$						
				MCI-2					
MCCL6	Statistics	39.9	.48	6.7	13.8	11.7	16.3	**39.9**	11.7
MCCL7	Science	54.8	.41	9.2	21.2	**54.8**	6.7	7.1	1.1
MCCL8	Statistics	59.0*	.40	14.1	8.8	8.1	**59.0**	6.7	3.2
MCRO6	Statistics	28.3*	.45	16.0	23.1	12.8	15.6	4.3	**28.4**[b]
MCRO7	Math	42.1	.43	9.5	33.6	**42.1**	9.2	3.9	1.8
MCRO8	Math	68.2	.44	2.1	9.2	15.2	**68.2**	3.2	2.1

Note. For MC items, the correct answer is underlined and the most common response is in bold.

[a] The magnitude of the difficulty estimate and its interpretation have an inverse relationship. A smaller difficulty estimate indicates the TEI was more difficult than the MCI ($^+$) and a larger difficulty estimate indicates the TEI was easier than the MCI ($^-$).

[b] Percent correct responses and percent of participants choosing the correct response differ due to omitted responses.

$^+$TEI statistic was higher than MCI. $^-$TEI statistic was lower than MCI.

* $p < .004$.

Reliability and Unidimensionality. In the context of this study, the reliability of the instrument is not a statistic of primary interest because the data collection instrument was not built to assess a single set of constructs (i.e., it is not a test) and it was not intended to make claims about the ability of participants. Similarly, achieving unidimensionality was not a goal of the instrument building process, however it is a necessary assumption that must be met for the psychometric analyses that were

conducted. Both reliability and dimensionality were examined separately for each form of the test.

The reliability of the instrument was assessed by calculating Cronbach's Alpha (ltm package; Rizopoulos, 2018). Results showed Alpha = 0.787 for Form A and Alpha = 0.809 for Form B. Unidimensionality analyses were conducted using the NOHARM method (Fraser & McDonald, 1988, 2012) available through the sirt package (Robitzsch, 2020). Three models were compared assuming one, two, and three dimensions. Table 4.14 reports results for each model and test form for the Root Mean Square (RSM) statistic and Tanaka's goodness of fit index (GFI). All values indicate appropriate fit, RSM values were below the desirability threshold and GFI values were above 0.90.

Table 4.14

Dimensionality analysis

Statistic	1 Dimension	2 Dimensions	3 Dimensions
Form A			
RSM	.0107756	.0095305	.0083373
GFI	.9986573	.9989497	.9991962
Form B			
RSM	.0099918	.0083851	.0074237
GFI	.9991707	.9994160	.9995422

Note. RSM values are compared to a threshold given by $4 \times \frac{1}{\sqrt{n}}$ (where n is the sample size). This threshold had values of 0.2377 and 0.2268 for forms A and B respectively.

The 3-dimensional solution showed better RSM and GFI statistics for both forms than the 2-dimensional solution, and this solution in turn shows a better fit compared to the one-dimensional model. However, the values for both the RSM and GFI statistics are similar across models and the one-dimensional solution shows values that are well within acceptable values. Consequently, applying the principle of parsimony (Occam's razor; de

Ayala, 2013), these results do not support rejecting the one-dimensional model and unidimensionality of the instrument is assumed[6].

Item Response Theory

An item response theory 2-parameter logistic model was fit to estimate item parameters for difficulty (b) and discrimination (a) using the ltm package (Rizopoulos, 2018). All items were calibrated concurrently using all available observations. In addition to assuming unidimensionality based on the results of the previous section, the assumption of local independence across items was evaluated by calculating standardized LD-χ^2 statistics for all items (Chen & Thissen, 1997). Results indicated that the maximum value observed for this statistic was 3.2 and that there were no major threats to the local independence assumption (magnitudes larger than 10 are considered concerning; Cai et al., 2011).

The fit of the model was evaluated using the M_2 goodness-of-fit statistic (Maydeu-Olivares & Joe, 2005, 2006; Cai et al., 2006) and the Root Mean Square Error of Approximation (RMSEA; Cai et al., 2011). The values of both statistics indicate a significant lack of adequate fit ($M_2 = 31,070.40, df = 405, p < .001$ and $RMSEA = 0.36$). A 3-parameter logistic model (including a guessing parameter, c) was also fit to the data to evaluate relative model fit. Table 4.15 shows the log-likelihood estimate for both models, as well as the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978). A log-likelihood ratio test comparing these statistics was also conducted and the results are also shown in Table 4.15. Results suggest that the 3-parameter model does not fit the data significantly better than the 2-parameter

[6]Unidimensionality was corroborated by conducting log-likelihood ratio tests and modified parallel analyses, both conducted through methods available in the ltm package.

model. As discussed in Chapter 3, the guessing parameter is not considered informative when characterizing technology-enhanced items (Gifford, 2017; Huff & Sireci, 2011; Parshall & Harmes, 2014) and considering the lack of fit for both models, the 2-parameter model is preferred. The observed lack of fit can be attributed to the erratic behavior of some items (as will be shown in future sections). However, this lack of fit was not considered concerning because the purpose of this study is not to build a cohesive test for a single construct.

Table 4.15

Log-likelihood ratio test

Model	AIC	BIC	Log-likelihood	LRT	*df*	*p*
2-parameter	12,245.40	12,508.61	-6,062.70			
3-parameter	12,262.72	12,657.54	-6,041.36	42.67	30	.063

Note. LRT = Log-likelihood ratio test statistic.

IRT Item Difficulty and Discrimination

Item difficulty and discrimination estimates for all items (b and a respectively) based on the two-parameter model are shown in Table 4.16.

Results of the common block showed that item difficulties ranged between -0.68 (CL1) and 0.91 (RO2) while discrimination estimates ranged between 1.24 (RO2) and 2.47 (RO1). RO1 was not only the most discriminating item in this block but also the item with the largest discrimination parameter of all items in the instrument. Item characteristic curves for items in the common block are shown in Figure 4.1. Overall, items in the common block performed well, displaying moderate difficulties and appropriate discrimination estimates.

Table 4.16

2-PL IRT item parameter estimates

Item	*b*	*a*	Item	*b*	*a*
		Common Block			
MC1	0.14	1.29			
MC2	0.39	1.25			
CL1	-0.68	1.80			
CL2	-0.56	1.49			
RO1	-0.68	2.47			
RO2	0.91	1.24			
		Item Set 1			
TEI-1			MCI-1		
CL3	-1.06^{-}	0.87^{-}	MCCL3	-0.32	0.88
CL4	0.41^{+}	1.40^{+}	MCCL4	-1.00	1.03
CL5	0.35^{+}	1.62^{+}	MCCL5	0.10	1.56
RO3	-2.24^{+}	-0.46^{-}	MCRO3	-7.01	-0.13
RO4	-1,649.61^{-}	-0.001^{-}	MCRO4	-8.42	-0.15
RO5	-1.02^{-}	0.78^{-}	MCRO5	-0.27	0.80
		Item Set 2			
TEI-2			MCI-2		
CL6	0.37^{+}	1.83^{+}	MCCL6	0.31	1.60
CL7	-0.28^{-}	1.86^{+}	MCCL7	-0.26	1.19
CL8	-1.13^{-}	1.21^{+}	MCCL8	-0.47	1.07
RO6	0.54^{-}	1.05^{-}	MCRO6	0.80	1.54
RO7	-0.07^{-}	1.23^{-}	MCRO7	0.27	1.30
RO8	-1.14^{-}	1.08^{-}	MCRO8	-0.81	1.33

$^{+}$TEI statistic was higher than MCI. $^{-}$TEI statistic was lower than MCI.

Item parameters for item set 1 revealed item difficulties ranging from -1,649.61 to 0.41 for TEIs and from -8.42 to 0.10 for the multiple-choice counterparts. Meanwhile, discrimination parameters ranged between -0.46 and 1.62 for TEIs and between -0.15 and 1.56 for MCIs. Four items showed a negative discrimination parameter: RO3 and RO4 and their multiple-choice counterparts MCRO3 and MCRO4. However, the discrimination parameter for item RO4 is almost zero (-0.001). As it may be seen in the item characteristic curves for item set 1 shown in Figure 4.2, the curve corresponding to item

RO4 is almost flat. Moreover, item RO4 showed an extreme negative difficulty estimate (-1,649.61). Although the discrimination parameter of corresponding multiple-choice item (MCRO4) is also negative and fairly large in magnitude (-8.42) it is not as concerning as the estimate for the TEI version. Overall, for four item pairs in this block (CL3-MCCL3, CL4-MCCL4, CL5-MCCL5, and RO3-MCRO3) the TEI version showed a higher difficulty than its counterpart; for the remaining two item pairs, the opposite was true. Among the four item pairs with positive discrimination parameters, two showed a higher discrimination for TEIs (CL4-MCCL4 and CL5-MCCL5) while the remaining two showed higher discrimination for the multiple-choice versions.

Item parameters for the second item set showed similar patterns to those observed in item set 1 absent negative discrimination parameters. For five of the six items pairs the TEI version was more difficult than the MC counterpart with the exception of pair CL6-MCCL6. Item difficulties for item set 2 ranged from -1.14 to 0.54 for TEIs and between -0.81 and 0.80 for MCIs. Meanwhile, discrimination estimates ranged from 1.05 to 1.86 for TEIs and from 1.07 to 1.60 for MCIs. Finally, four TEIs showed higher discrimination than their corresponding multiple-choice versions (CL6, CL7, RO7, and CL8). Figure 4.3 shows the item characteristic curves for items in item set 2.

Figure 4.1

Item characteristic curves for the Common Block

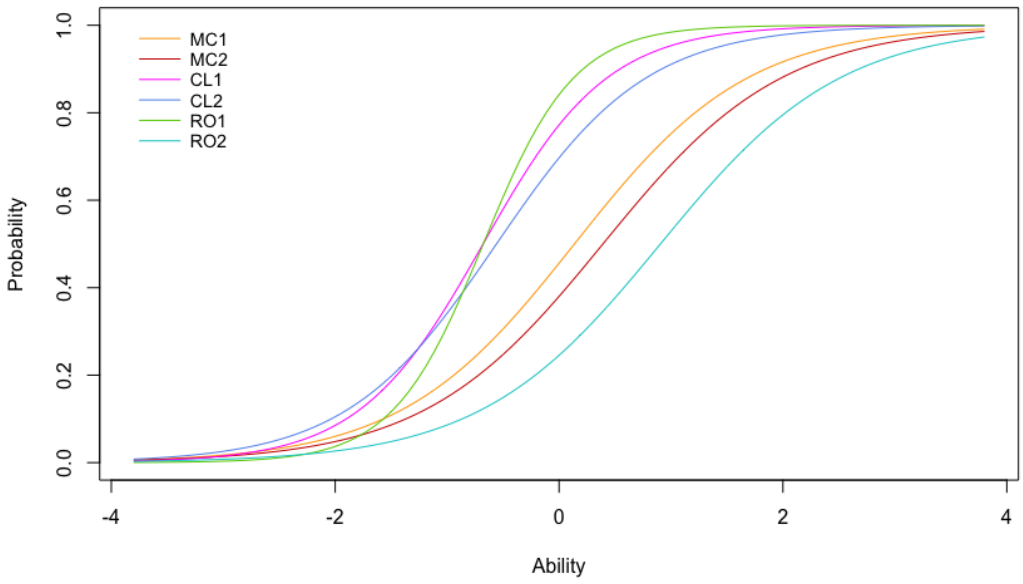


Figure 4.2

Item characteristic curves for Item Set 1

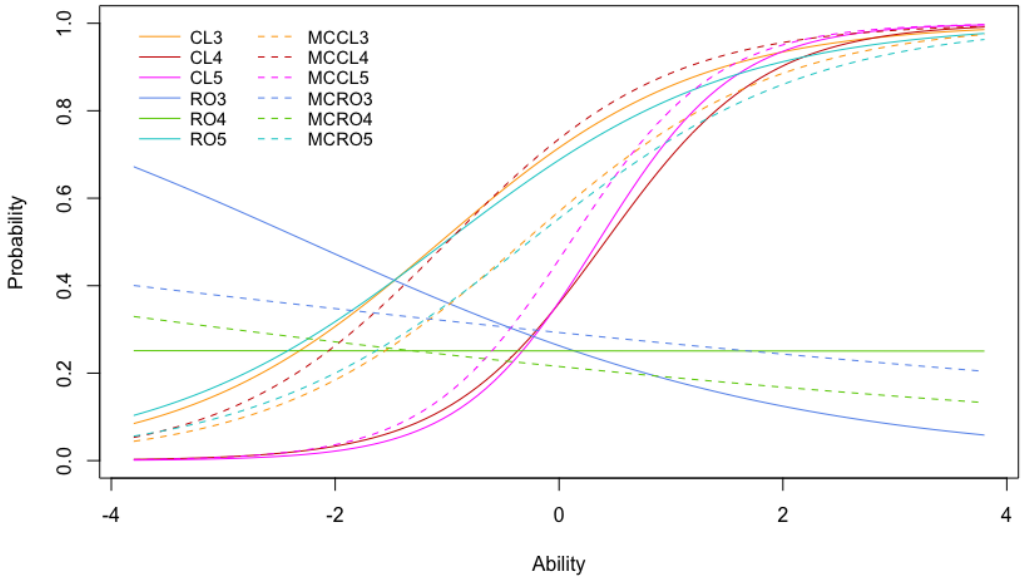
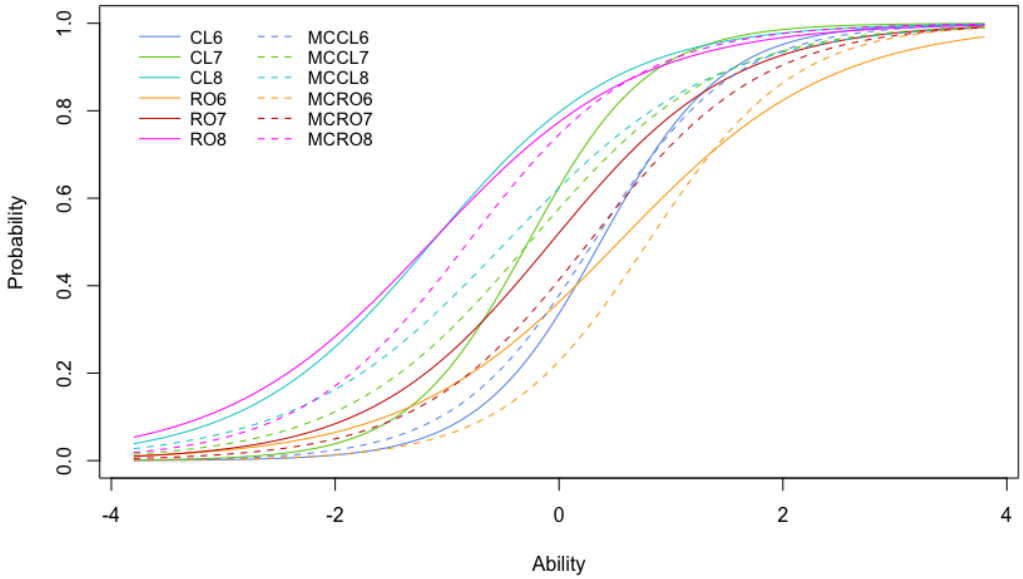


Figure 4.3

Item characteristic curves for Item Set 2



To compare item parameters of stem-equivalent items two types of graphical displays were used. Table 4.17 presents the difference between IRT parameters between stem-equivalent items in a pair as well as the respective standard error of the difference organized by response format. These differences were plotted in Figure 4.4 and a confidence interval was constructed around these differences based on the standard error of the difference and the 95% critical value of the z distribution for 0.004 (significance level corrected for multiple comparisons). Items RO3 and RO4 are not shown due to the magnitudes of their differences in difficulty. As described in Chapter 3, the TIMSS & PIRLS International Study Center employs this approach to identify trend items that present parameter drift across modes or between consecutive administrations, flagging differences larger than 2 logits as concerning (Fishbein et al., 2020). Results show that

only items RO3 and RO4 presented significant differences in their difficulty parameters across item formats.

Table 4.17

Differences in item parameter estimates between stem-equivalent items

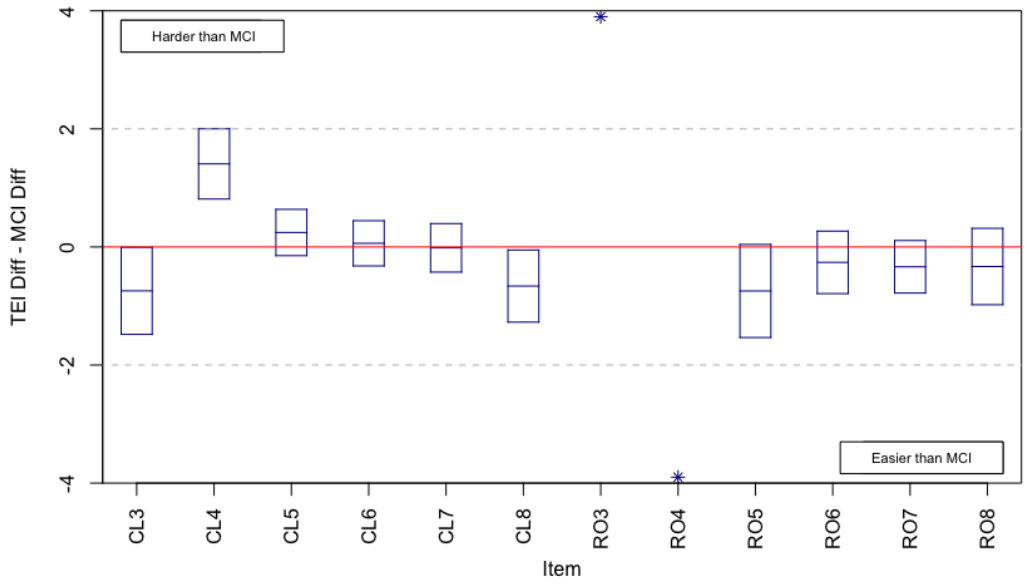
Item Pair	Difficulty		Discrimination	
	Difference	Std. Error	Difference	Std. Error
		Classification		
CL3-MCCL3	-0.74	0.28	-0.01	0.25
CL4-MCCL4	1.41	0.23	0.37	0.29
CL5-MCCL5	0.25	0.15	0.06	0.35
CL6-MCCL6	0.06	0.15	0.23	0.37
CL7-MCCL7	-0.02	0.16	0.67	0.34
CL8-MCCL8	-0.66	0.23	0.13	0.29
		Rank-Ordering		
RO3-MCRO3	4.77	7.63	-0.33	0.21
RO4-MCRO4	-1,641.19	384,405.54	0.15	0.22
RO5-MCRO5	-0.75	0.30	-0.03	0.23
RO6-MCRO6	-0.26	0.20	-0.48	0.32
RO7-MCRO7	-0.34	0.17	-0.07	0.29
RO8-MCRO8	-0.33	0.25	-0.25	0.30

The graphical display shown in Figure 4.4 cannot be replicated for discrimination parameters as these cannot be interpreted in the logits scale. Consequently, a second graphical display was employed following Gaertner and Briggs's (2009) "3-sigma IRT" approach to identify item parameter drift. In this graphical display item parameters for each pair of stem-equivalent items are plotted against each other (Figure 4.5). The standard deviation line (SD line) is graphed as the line of best fit and a confidence region is plotted around this line that represents three times the standard deviation of the perpendicular distances between each point and the SD line. Points beyond this region are considered to indicate significant differences in their item parameters. The diagonal line

$y = x$ was also plotted in this graphical display as a reference line. Points above this line indicate that TEIs were easier than their MC counterparts, while points below this line indicate TEIs were harder than the corresponding MCIs. Items RO3 and RO4 were not included in Figure 4.5 nor were they included as part of the calculations due to their extreme values. Results indicate that items CL4 and MCCL4 showed significant difference in their difficulty parameter. This graphical display was replicated for discrimination parameters and is shown in Figure 4.6. In this display, points above the $y = x$ reference line indicate the MCI discriminated better than the TEI counterpart, while points below the line indicate the opposite. Results indicate that all items had similar discrimination parameters with the exception of items RO6 and MCRO6.

Figure 4.4

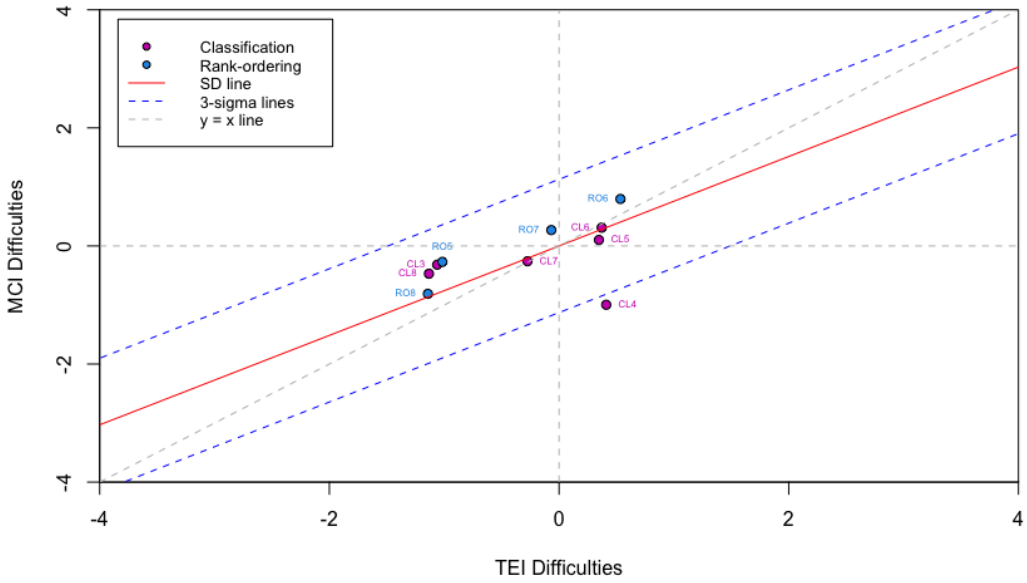
Differences in difficulty parameters across stem-equivalent items



Note. Items RO3 and RO4 are not shown due to the magnitudes of their differences and their confidence intervals: 4.77 [-15.4, 24.9] and -1,641.19 [-1,015,816, 1,012,533] respectively.

Figure 4.5

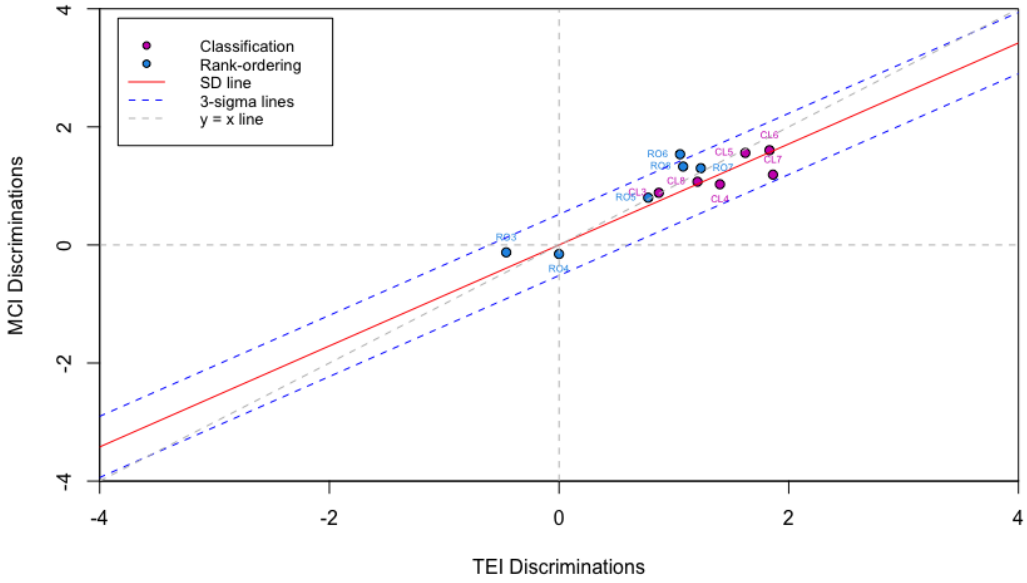
Item difficulty parameters (b) of stem-equivalent items across formats



Note. Items RO3 and RO4 are not shown due to the magnitudes of their difficulty parameters and were not included in the calculation of the SD line or its confidence region.

Figure 4.6

Item discrimination parameters (a) of stem-equivalent items across formats



IRT Information and Relative Efficiency

This section presents analyses based on item information estimates provided by the 2-parameter model used. First, item information curves for all items are shown, followed by relative efficiency curves (Lord, 1980), average expected information estimates, and measurement efficiency statistics (Donoghue, 1994; Jodoin, 2003).

Item information curves for all items are presented in Figures 4.7, 4.8, and 4.9 for the common block, item set 1, and item set 2 respectively (note that for clarity Figure 4.7, and Figures 4.8 and 4.9 have different vertical scales). Most items achieve maximum information between theta values of -2 and 2. Item RO1 shows the highest maximum information value at about 1.5 among all items, a direct consequence of its high discrimination. Items RO3 and RO4 and their multiple-choice counterparts provide little to no information across the totality of the ability spectrum (flat curves) due to their negative discrimination parameters.

Figures 4.8 and 4.9 showcase how TEIs may provide more information than their multiple-choice stem-equivalent counterparts in certain regions of the ability continuum while providing less information in other regions. In item set 1, some item pairs show almost no difference between their information curves (e.g., CL3-MCCL3 and RO5-MCRO5) while others show more visible differences (CL4-MCCL4 and CL5-MCCL5). The difference between the information curves for CL4 and MCCL4 is the starkest difference between two items observed among all 12 pairs (i.e., both item sets). While item CL4 attains maximum information of about 0.5 around 0.4 logits, item MCCL4 provides maximum information of 0.3 at -1 logits. This indicates that CL4 provides more information than MCCL4 for more than half of the ability continuum ($\theta > -0.5$).

For items in item set 2 item difficulty estimates were fairly similar across response formats with differences ranging between -0.66 and 0.06 (Table 4.17). Consequently most item information curves of stem-equivalent items in this item set were centered around similar locations with the exception of item pair CL8-MCCL8, for which the TEI attained its maximum at about -1 logits and the MCI attained its maximum closer to zero. Despite being fairly well aligned, differences in the maximum information attained were observed. Items CL6 and CL7 achieved a higher information maximum compared to their multiple-choice counterparts (differences of 0.2 and 0.5 respectively). In contrast, the opposite was observed for items RO6, RO7, and RO8.

Figure 4.7

Item information curves for the Common Block

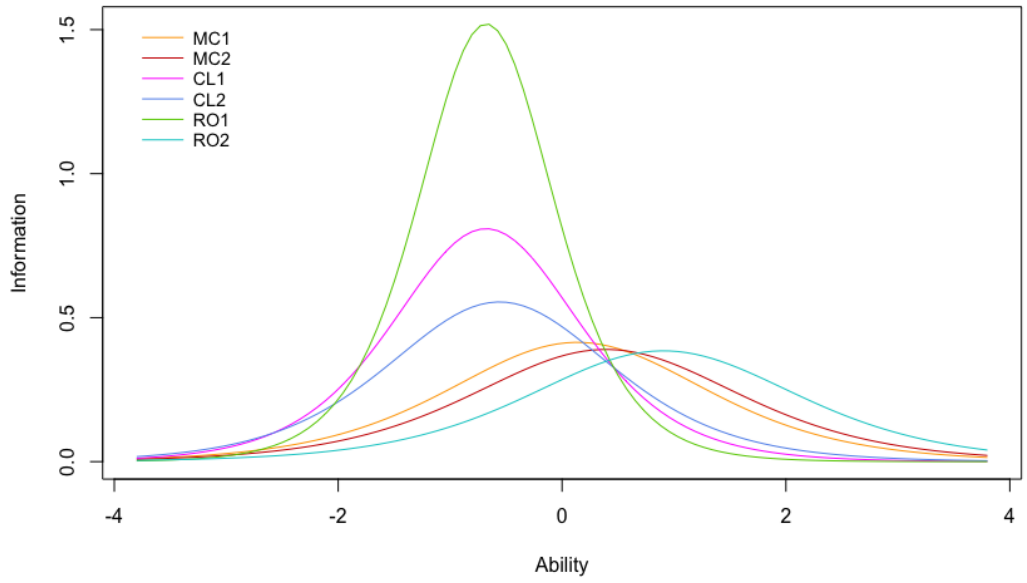
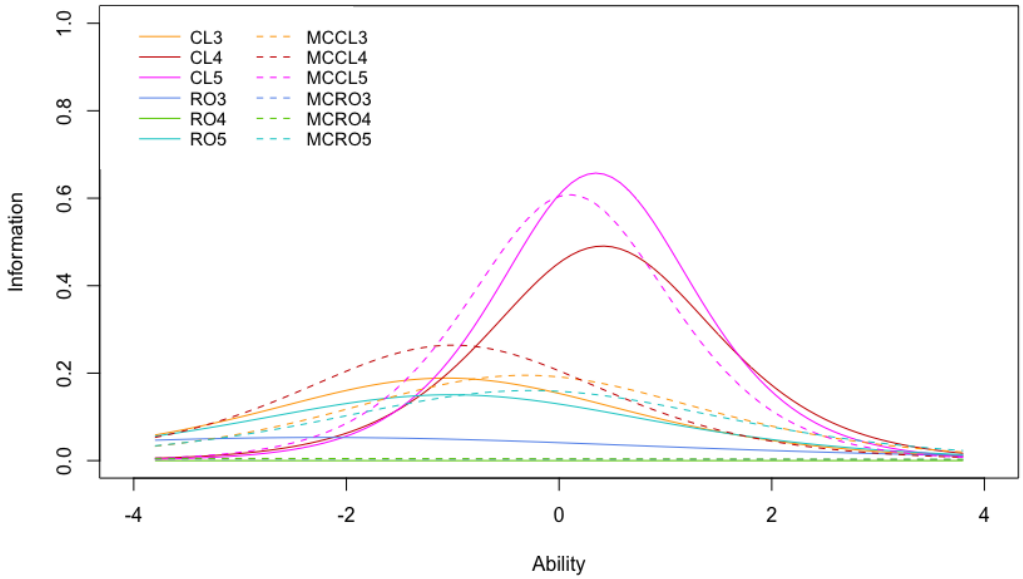


Figure 4.8

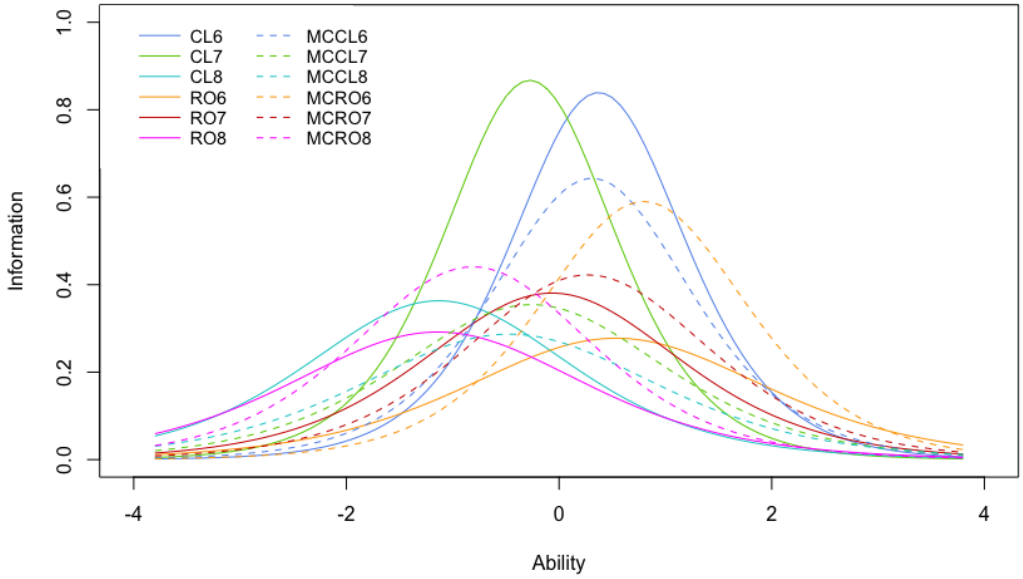
Item information curves for Item Set 1



Note. The curve for item MCRO3 overlaps with the curve for item MCRO4.

Figure 4.9

Item information curves for Item Set 2



Comparing information curves visually is not straightforward because the relationship between two curves changes at different regions of the ability continuum. An alternative approach is to calculate a relative efficiency function (Lord, 1980). Relative efficiency curves were calculated for each stem-equivalent item pair as a ratio of TEI information to MCI information ($RE\{TEI, MCI\}$). Figures 4.10 and 4.11 show relative efficiency curves for classification items (CL3-CL8) and rank-ordering items (RO3-RO8) respectively. Relative efficiency curves are interpreted by comparing them to a horizontal reference line equal to 1 (black dotted line). Regions where $RE\{TEI, MCI\}$ is above 1, indicate the TEI is more efficient than its multiple-choice counterpart in that range of ability level (i.e., the TEI provides more information than the MCI in that region). Conversely, regions where $RE\{TEI, MCI\}$ is below 1, indicate the multiple-choice version is more efficient than the TEI at that range of ability level.

The results shown in Figure 4.10 indicate that the relative efficiency of classification items varied greatly. No classification TEI provided more information than its multiple-choice counterpart throughout the whole range of ability; instead, TEIs were more efficient in specific regions. While items CL3 and CL8 provided more information than their MC counterparts for low-ability participants ($\theta < -1$), items CL4 and CL5 provided more information for participants with average to high ability ($\theta > 0$). Items CL6 and CL7 provided more information than their corresponding MCIs for average ability participants ($-1 < \theta < 2$ and $-2 < \theta < 1$ respectively).

Relative efficiency curves for rank-ordering items were more erratic than those of classification items. Item RO3 appeared to be more efficient than MCRO3 while item RO4 appeared to be less efficient than MCRO4 throughout the whole spectrum of ability.

However, these two items had almost flat information curves due to their negative discrimination, thus making these results inconclusive. The scale of the vertical axis in Figure 4.11 impedes clear comparisons of relative efficiency among all other items. Figure 4.12 provides a focused region of Figure 4.11. Items RO5 and RO7 were more efficient than their MC counterparts for low-ability participants ($\theta < -1$). Items RO6 and RO8 were more efficient than their corresponding multiple-choice items in the extremes of the ability continuum, while being less efficient at the middle of the ability continuum ($-1 < \theta < 3$ and $-2 < \theta < 2$ respectively).

Figure 4.10

Relative efficiency curves of Classification TEIs vs. stem-equivalent MCIs

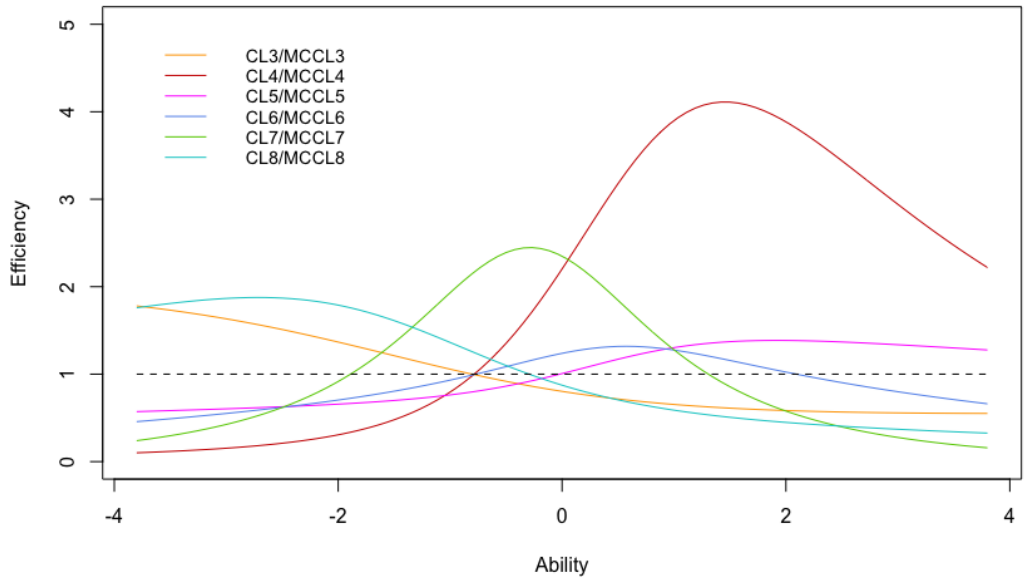


Figure 4.11

Relative efficiency curves of Rank-ordering TEIs vs. stem-equivalent MCIs

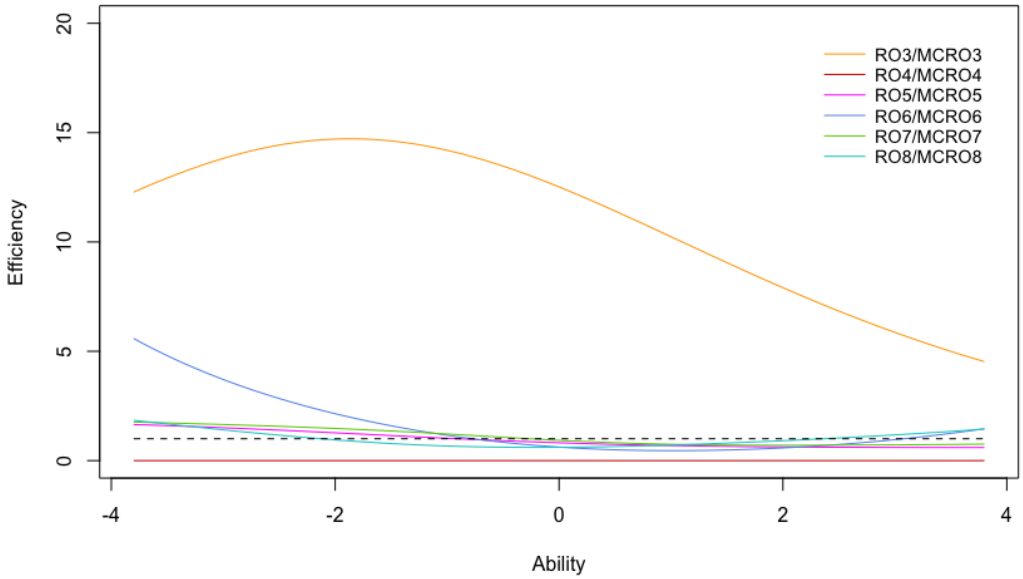
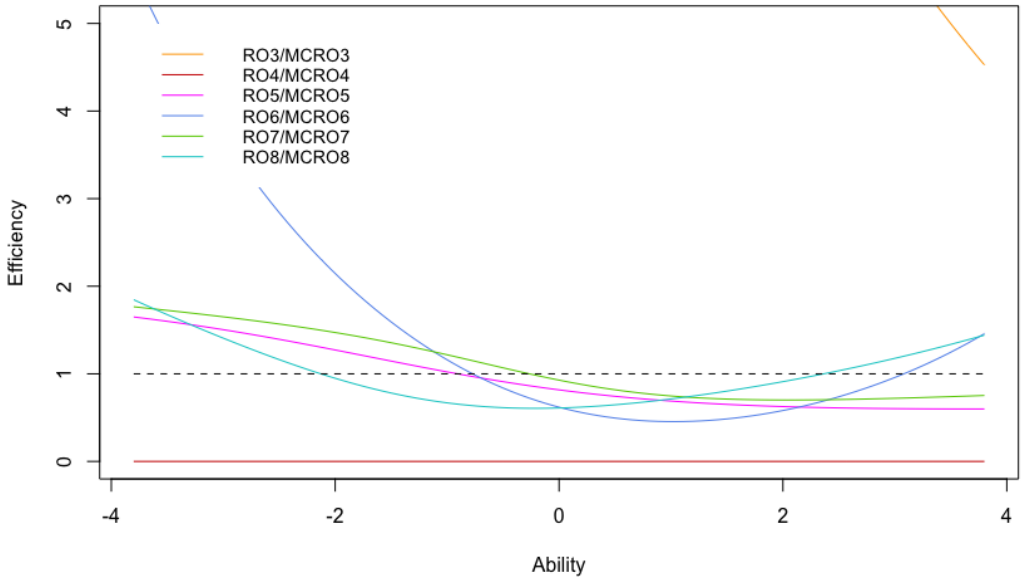


Figure 4.12

Relative efficiency curves of Rank-ordering TEIs vs. stem-equivalent MCIs (region)



In contrast to Lord's (1980) relative efficiency curves, Jodoin's (2003) approach produces a point-estimate based on the ratio between the average expected information for each item and the median time spent by participants on that item (measurement efficiency). Average expected information was calculated as the average item information evaluated at each participants' ability level (catIrt package, Nydick, 2015).

Table 4.18 shows the estimate of the average expected information —$E(I_j)$, the median time, and the measurement efficiency (expected information per minute; $E(I_j)$/min) for each item. As the purpose of this study is to compare stem-equivalent items, this table only shows results for items in item sets 1 and 2 organized by item format. For five of the item pairs the TEI version provided more expected information than its multiple-choice version. In particular, this was true for four of the six classification item pairs but only one rank-ordering item pair. These results vary slightly when examining measurement efficiency. Three classification TEIs provided more expected information per minute than their MC counterparts while this relationship was observed for only one rank-ordering item (RO3). These results diverge slightly from previous studies (Donoghue, 1994; Jodoin, 2003; Wainer & Thissen, 1993) in which multiple-choice items often provided more information per minute. However, in contrast with previous studies, the multiple-choice items employed in this study often required the same or more time than the corresponding TEIs. Table 4.18 also shows the mean of the average expected information and measurement efficiency for each item type. These results indicate that drag-and-drop classification items provided more average expected information and higher measurement efficiency than their multiple-choice counterparts. However, this relationship did not hold for rank-ordering drag-and-drop items.

Table 4.18

Item expected information and expected information per minute

Item	$E(I_j)$	Mdn Time	$E(I_j)$/min	Item	$E(I_j)$	Mdn Time	$E(I_j)$/min
			Classification				
CL3	.143^{-}	0.81^{-}	0.177^{-}	MCCL3	.168	0.87	0.193
CL4	.352^{+}	1.51^{+}	0.233^{+}	MCCL4	.189	1.39	0.136
CL5	.443^{+}	1.24^{+}	0.357^{-}	MCCL5	.429	1.20	0.358
CL6	.525^{+}	1.26^{-}	0.417^{+}	MCCL6	.438	1.33	0.330
CL7	.547^{+}	0.48^{-}	1.140^{+}	MCCL7	.278	0.68	0.410
CL8	.227^{-}	0.73^{+}	0.310^{-}	MCCL8	.228	0.72	0.317
Means	0.373		0.439		0.288		0.290
			Rank-Ordering				
RO3	.040^{+}	0.68^{-}	0.059^{+}	MCRO3	.003	0.74	0.004
RO4	$<$ .001^{-}	0.66^{-}	$<$ .001^{-}	MCRO4	.004	0.77	0.005
RO5	.121^{-}	0.39^{-}	0.309^{-}	MCRO5	.141	0.45	0.314
RO6	.219^{-}	0.81^{-}	0.271^{-}	MCRO6	.361	0.87	0.415
RO7	.299^{-}	0.52^{+}	0.574^{-}	MCRO7	.320	0.46	0.695
RO8	.194^{-}	0.91^{+}	0.213^{-}	MCRO8	.294	0.82	0.358
Means	0.145		0.238		0.187		0.299

$^{+}$TEI statistic was higher than MCI. $^{-}$TEI statistic was lower than MCI.

To simplify comparisons across items of a stem-equivalent pair, ratios of expected information and expected information per minute were calculated for each item pair (TEI:MCI) and results are shown in Table 4.19 grouped according to the response format of the TEI (classification or rank-ordering). Results show that TEIs provided more expected information than their multiple-choice counterparts for four of the six classification item pairs, and greater measurement efficiency for three of these pairs. Item CL7 provided over twice as much expected information per minute than its counterpart MCCL7. In contrast, for rank-ordering items the opposite was observed. Disregarding the statistics for item pairs RO3-MCRO3 and RO4-MCRO4 due to their erratic psychometric behavior, it is reasonable to conclude that rank-ordering items provided less expected

information and less measurement efficiency compared to their multiple-choice counterparts.

Table 4.19

Relative expected information and relative measurement efficiency for each item pair

Item Pair	Relative Expected Information Ratio of $E(I_j)$	Relative Measurement Efficiency Ratio of $E(I_j)$/min
	Classification	
CL3:MCCL3	0.854	0.917
CL4:MCCL4	1.860	1.712
CL5:MCCL5	1.031	0.998
CL6:MCCL6	1.198	1.264
CL7:MCCL7	1.965	2.783
CL8:MCCL8	0.993	0.979
	Rank-ordering	
RO3:MCRO3	12.355	13.446
RO4:MCRO4	$< .001$	$< .001$
RO5:MCRO5	0.853	0.985
RO6:MCRO6	0.607	0.652
RO7:MCRO7	0.934	0.826
RO8:MCRO8	0.659	0.593

TEI Utility Ratings

A panel of three graduate students in the Measurement, Evaluation, Statistics, and Assessment (MESA) program at Boston College employed the TEI Utility Framework (Russell, 2016; Russell & Moncaleano 2019) to evaluate the construct fidelity of each TEI included in the instrument and the usability of each TEI response format (classification and rank-ordering). A total of 16 TEIs were reviewed by the panelists, four items from the common block and all items in blocks TEI-1 and TEI-2. Initial responses showed agreement on the construct fidelity ratings for five items, while 11 items showed disagreement. A summary of initial panelist ratings is shown in Table 4.20. In the five instances of agreement, all panelists rated the construct fidelity to be moderate (i.e.,

authentic context and inauthentic actions). The most common pattern of ratings occurred when two panelists rated the TEI to have high construct fidelity (i.e., authentic context and actions) while a third panelist rated the TEI to have moderate fidelity.

Table 4.20

Summary of initial TEI construct fidelity ratings

Panelist Ratings	Number of items
Moderate — Moderate — Moderate	5
High — High — Moderate	7
High — Moderate — Moderate	2
High — Moderate — None	2

After the initial ratings were completed, the panelists reconvened to discuss their ratings explaining to other panelists their reasoning with the aim to achieve a consensus. After this round of discussions, all panelists agreed that all TEIs had moderate construct fidelity. Panelists argued that all contexts presented in the items were authentic but that the actions test-takers were required to take to produce a response were inauthentic as they did not resemble how the assessed constructs are employed in the real world. Panelists often felt most of these items would be better served with other response formats, such as multiple-choice or open-ended prompts.

Panelists also rated the usability of each TEI response format as a class. Panelists' initial ratings showed total agreement that both the classification drag-and-drop and rank-ordering drag-and-drop response formats showed high usability. Usability was not rated on an item-by-item basis because the instrument-delivery platform (Qualtrics) only provides a single way of employing these response interactions.

Summary

Based on the results described in this chapter, this section summarizes the results in relation to the three research questions of the study.

RQ1: How do the psychometric characteristics of commonly employed TEI drag-and-drop formats (classification and rank-ordering) compare to stem-equivalent multiple-choice items?

Table 4.21 presents a summary of how the TEI version of an item compared to its stem-equivalent multiple-choice version for each of the estimated item characteristics discussed in this chapter: CTT difficulty and discrimination, IRT difficulty and discrimination, information curves, and efficiency. For each item statistic, this table shows whether the TEI version had a higher or lower statistic than its MC counterpart. Comparisons between information curves were characterized in three regions: low-performing participants ($\theta < -1$), average-performing participants ($-1 \leq \theta \leq 1$), and high-performing participants ($\theta > 1$). Lord's relative efficiency ratio is not included in this table as it is redundant with comparisons between information curves. Recall that comparing efficiency statistics ($E(I_j)$ and $E(I_j)$/min) within a stem-equivalent pair is equivalent to comparing the corresponding ratios to 1 (e.g., $E(I_j)$ for a TEI is higher than $E(I_j)$ for its MCI counterpart if and only if $E(I_{TEI})/E(I_{MCI}) > 1$). Jodoin (2003) considered relative efficiency values (ratios of expected information per minute) larger than 2 to be meaningful, this criterion was extended to ratios of expected information. Items that showed a magnitude larger than 2 in either of these ratios are flagged in 4.21.

Table 4.21

Summary of comparisons between stem-equivalent TEIs and MCIs across item characteristics

Item	CTT		IRT		Information			Efficiency	
	Diff[a]	Disc	b	a	$\theta < -1$	$-1 < \theta < 1$	$1 < \theta$	$E(I_j)$	$E(I_j)$/min
					Classification				
CL3	-*	-	-	-	+	-	-	-	-
CL4	+*	+	+*	+	-	+	+	+	+
CL5	+	-	+	+	-	∼	+	+	-
CL6	-	+	+	+	-	+	∼	+	+
CL7	-	+	-	+	∼	+	-	+	+†
CL8	-*	-	-	+	+	∼	-	-	-
					Rank-ordering				
RO3	+	-	+*	-	+	+	+	+†	+†
RO4	-	+	-*	+	-	-	-	-‡	-‡
RO5	-	-	-	-	+	-	-	-	-
RO6	-*	-	-	-*	+	-	∼	-	-
RO7	-	+	-	-	+	∼	-	-	-
RO8	-	-	-	-	∼	-	∼	-	-

[a] The magnitude of the CTT difficulty estimate and its interpretation have an inverse relationship. For this statistic, + indicates the TEI is more difficult than the MCI, not that the value of the estimate is larger. Similarly, - indicates the TEI is easier than the MCI, not that the value of the estimate is smaller.

+ TEI statistic was higher than MCI. - TEI statistic was lower than MCI. ∼ TEI information curve appeared both higher and lower than the MCI curve within the specified region.

*Difference between TEI and MCI statistic was considered statistically significant.

† Ratio across formats (TEI:MCI) was meaningfully large (i.e., > 2).

‡ Ratio across formats (TEI:MCI) was meaningfully small (i.e., < 0.5).

Classification items. CTT and IRT parameters showed a moderate degree of agreement. Items CL3, CL7, and CL8 appeared to be easier than MCCL3, MCCL7, and MCCL8 according to both the CTT p-value and the IRT difficulty parameter, while the opposite was observed for items CL4 and CL5. Item CL6 appeared to be easier according to the CTT p-value but harder based on the IRT parameter estimate. However, only for item (CL4) was the difference in difficulty significant for both CTT and IRT parameters.

Items CL4, CL6, and CL7 provided better discrimination according to both CTT and IRT parameters while item CL3 had lower discrimination estimates. The remaining items showed discrepancies between the CTT and IRT discrimination estimates. All item information curve comparisons were different for each item; however it appears that generally TEIs provided more information than their multiple-choice counterparts for average-ability participants. Finally, four out of the six classification items provided more expected information than their stem-equivalent pairs, and three of these items showed higher measurement efficiency (expected information per minute). Item CL7 was the only item to provide more than double the expected information per minute than its multiple-choice counterpart.

Rank-ordering items. For these item pairs, there was almost perfect alignment between CTT and IRT statistics (both difficulty and discrimination). Only item RO3 appeared to be more difficult than its counterpart while the remaining five items appeared to be easier. Four of the six TEIs had lower CTT and IRT discrimination estimates, while item RO4 had higher discrimination. Only for item (RO7) did the CTT and IRT discrimination parameters diverge. Item RO3 provided more information than its counterpart throughout all of the ability continuum, while the opposite was observed for item RO4. However, keep in mind that both of these items presented negative discrimination parameters, thus their information curves were flat. The comparisons between information curves for the remaining four items showed that rank-ordering TEIs appeared to provide more information for low-ability participants. Finally, five out of the six rank-ordering TEIs provided less average expected information and measurement efficiency than their stem-equivalent counterparts.

In summary, few items showed significant differences in their item parameters. In particular, only four items showed important differences in item parameters according to IRT estimates (and two of these items had erratic psychometric behaviors —RO3 and RO4) suggesting that keeping the stem equivalent across item formats helps to control for these item characteristics. Although not significantly different, classification TEIs appeared to be more discriminating compared to their MC pairs than rank-ordering TEIs compared with their respective MC counterparts. Moreover, classification items also appear to provide more expected information and expected information per minute than their MC counterparts compared to rank-ordering item pairs.

RQ2: What is the relationship between the utility of TEI drag-and-drop formats (classification and rank-ordering) and their psychometric item characteristics?

All technology-enhanced items used in this study were rated by the panel as having moderate construct fidelity according to the TEI Utility Framework (Russell, 2016). This indicates that the TEIs used in this study presented authentic contexts to participants to demonstrate their knowledge, but the response interactions used were not authentic. In other words, panelists believed that classification and rank-ordering drag-and-drop interactions do not align with the way targeted constructs are observed outside a testing environment. The panel also rated both drag-and-drop response formats (classification and rank-ordering) of high usability. Unfortunately, the lack of variability of the TEI utility ratings limit the capacity to identify relationships between them and psychometric item characteristics.

RQ3: How can TEI psychometric properties and utility ratings be combined to develop a standardized protocol to judge the comparative measurement value of TEIs relative to stem-equivalent MC items?

The results presented in this chapter showcase how TEIs and stem-equivalent multiple-choice items compare to each other based on multiple psychometric properties. Moreover, the construct fidelity and usability ratings obtained from the panel allow judging the quality of the TEIs used. However, combining these psychometric properties and utility ratings to develop a decision protocol (and answer this research question) is primarily a judgment-based task that draws on the empirical evidence presented in this chapter. Consequently, this research question is explored fully in the following chapter.

Chapter 5 - Discussion

This dissertation aims to develop a protocol to guide test developers as they make judgments about the measurement value of technology-enhanced items when compared to stem-equivalent multiple-choice items. The protocol also applies this judgment regarding comparative measurement to guide decisions about when to give preference to a TEI or MCI format. To inform the development of the protocol, the following research questions were explored:

1. How do the psychometric characteristics of commonly employed TEI drag-and-drop formats (classification and rank-ordering) compare to stem-equivalent multiple-choice items? (RQ1)

2. What is the relationship between the utility of TEI drag-and-drop formats (classification and rank-ordering) and their psychometric item characteristics? (RQ2)

3. How can TEI psychometric properties and utility ratings be combined to develop a standardized protocol to judge the comparative measurement value of TEIs relative to stem-equivalent MC items? (RQ3)

This chapter begins with a summary of the findings presented in Chapter 4 that address the first two research questions. The third research question is then addressed by presenting the comparative measurement value protocol and then demonstrating its application using the items administered as part of this study. This chapter also discusses implications of this study for the field of educational measurement, limitations of the study and its results, and future steps for this line of research.

Summary of Findings

The empirical evidence presented in Chapter 4 that addresses RQ1 indicates that CTT and IRT item parameters were similar for each stem-equivalent pair of items. A larger increase in discrimination parameters was observed for classification TEIs compared to their multiple-choice counterparts than was observed for rank-ordering items. However, these differences were not significant. Classification items also provided more expected information and measurement efficiency (expected information per minute) than their MC counterparts compared to rank-ordering item pairs.

Results of RQ2 were less informative because the panel of graduate students rated all of the TEIs used for this study as having moderate construct fidelity. These ratings indicate that all technology-enhanced items used provided an authentic context but the response interactions were inauthentic in relation to contexts outside of testing. Although moderate construct fidelity is acceptable, the lack of variability in the ratings resulting from the panels work prevented further study of the relationships between construct fidelity and psychometric properties. The panel also rated the two TEI drag-and-drop formats (classification and rank-ordering) as having high usability. The third component of TEI utility (accessibility) was not rated as the test-delivery platform (Qualtrics) used for this study did not provide any accessibility accommodations.

Combining the psychometric properties and utility ratings to construct a decision protocol (addressing RQ3) is primarily a judgment-based task that draws on the empirical evidence presented in the previous chapter. This third research question is explored at length in the following section.

The Comparative Measurement Value Protocol

This section presents the development of the comparative measurement protocol, the application of this protocol to the stem-equivalent item pairs used in this study, and instructions for future use.

Selection of Indicators

As described in Chapters 3 and 4, nine characteristics were estimated for each item to address research questions 1 and 2, namely, (a) CTT difficulty, (b) CTT discrimination, (c) IRT difficulty, (d) IRT discrimination, (e) IRT item information, (f) expected information, (g) expected information per unit of time, (h) TEI construct fidelity, and (i) TEI usability. These characteristics were evaluated to identify the indicators that are most informative for judging the measurement value of TEIs relative to stem-equivalent multiple-choice items.

The summary presented in Table 4.21 indicates the results of comparisons between stem-equivalent items showed agreement between CTT and IRT statistics about 80% of the time (20/24 cells), indicating the information provided by CTT and IRT statistics was redundant. Differences in difficulty estimates across stem-equivalent items are only meaningful if these differences are significant. In other words, it is of no concern whether the TEI version is more or less difficult than the multiple-choice version, but rather whether the difference is large enough to be significant. Similarly, differences in discrimination, although interpretable at face value, may be small and inconsequential making it more valuable to focus again only on differences that are significant. CTT statistics have the benefit of allowing t tests to determine the significance of the difference between statistics across item formats. In contrast, comparing IRT parameters requires

multi-step methods analogous to identifying parameter drift. Even though CTT estimates allow for easier evaluation of statistical significance than IRT estimates, the latter are sample-free and are often considered more reliable. Consequently, despite their added complexity, IRT parameters are preferred as indicators for the comparative measurement value protocol.

Comparing item information is challenging because the relationship between two item information curves changes across the ability continuum. Even though Lord's (1980) relative efficiency ratio function may be used to simplify these comparisons, this curve and its interpretation also vary across the ability continuum. In contrast, expected information and measurement efficiency (expected information per minute) are point estimates that allow unambiguous direct comparisons between stem-equivalent items. Recall that expected information is used when estimating measurement efficiency and, according to the results shown in Table 4.21, these two characteristics provide redundant information about stem-equivalent items about 90% of the time. Although previous research indicates that multiple-choice items often require less time to answer than TEIs, this study found item pairs for which the opposite occurred, which highlights the importance of using response time to evaluate comparative measurement value. Given that measurement efficiency considers both expected information and response time and yields a point estimate that does not require human judgment, measurement efficiency is included in the decision protocol as an indicator.

Finally, TEI utility ratings provide useful information about the extent to which a TEI improves construct representation. Although usability and accessibility are important properties of TEIs, construct fidelity indicates whether a TEI provides a more authentic

context to test-takers than traditional multiple-choice items. Therefore, ratings of construct fidelity are deemed a vital indicator for the construction of this protocol.

In conclusion, IRT difficulty and discrimination item parameters, expected information per minute (measurement efficiency), and construct fidelity were selected as the most informative characteristics for judging the comparative measurement value of a TEI in relation to a stem-equivalent multiple-choice item. A protocol that uses these four characteristics to judge comparative measurement is described in the next section.

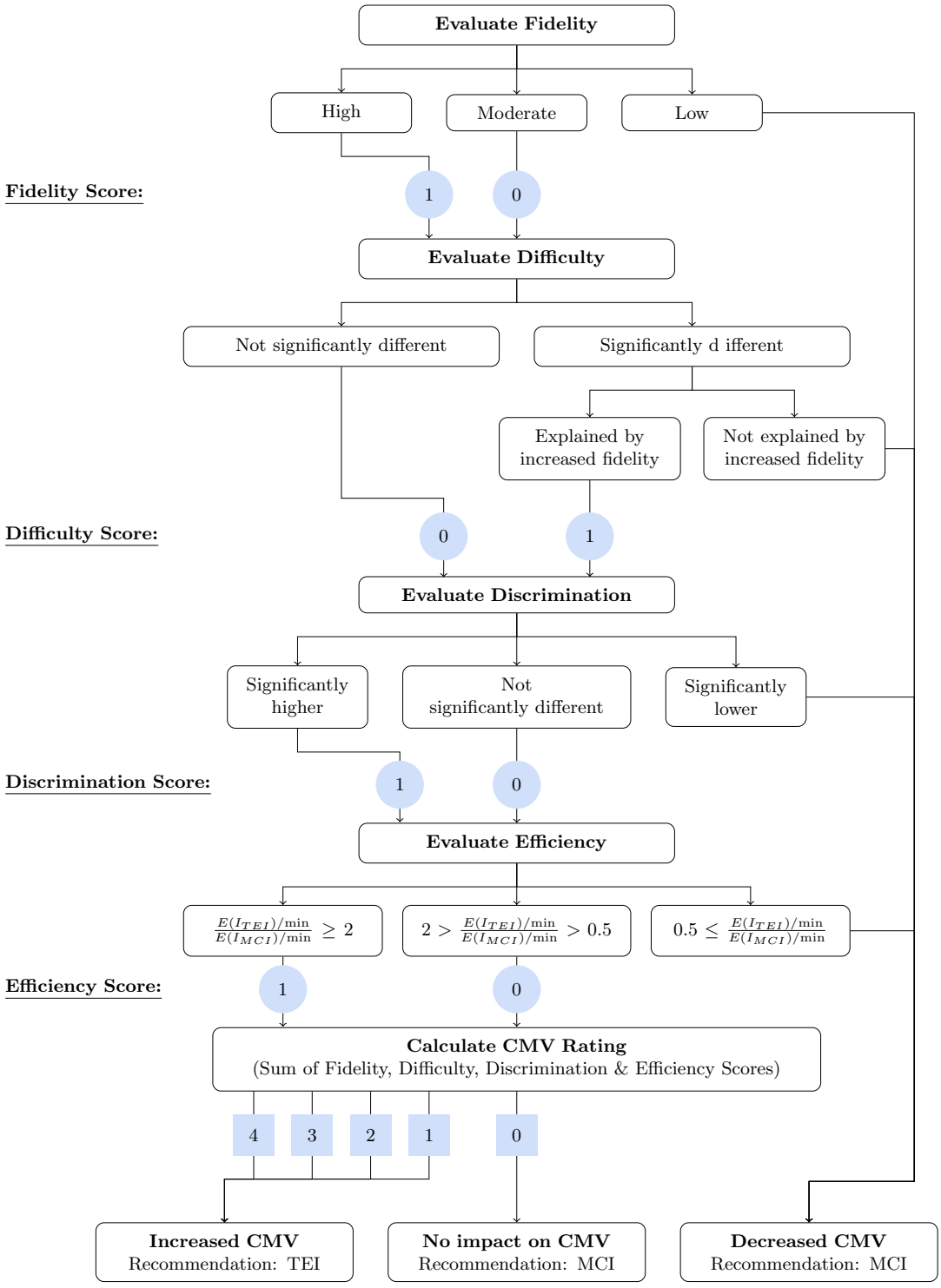
Development of the Comparative Measurement Value Protocol

Analyses presented in Chapter 4 and summarized above revealed that three quantitative item characteristics, namely IRT difficulty, IRT discrimination, and measurement efficiency, and one qualitative item characteristic, namely construct fidelity, were non-redundant, informative indicators for comparing stem-equivalent pairs of multiple-choice and technology-enhanced items. This section discusses how these characteristics are combined to produce a judgment of comparative measurement value (CMV) and a subsequent recommendation regarding the preferred item format. The resulting protocol is shown in Figure 5.1 as a four-step decision tree.

The four selected indicators were organized into a hierarchical decision tree that guides users through four steps: (a) evaluating construct fidelity, (b) evaluating difficulty, (c) evaluating discrimination, and (d) evaluating efficiency. As is described in detail next, at each step, the first consideration focuses on determining whether an egregious impact on measurement value was produced by the TEI. If so, use of the multiple-choice format is immediately recommended. If not, then additional criteria are considered to determine whether the use of the TEI produced a positive or neutral impact on measurement value.

Figure 5.1

The Comparative Measurement Value Protocol



Note. A significant difference in difficulty for a TEI judged to have high construct fidelity is assumed to be construct-relevant (i.e., explained by increased fidelity).

Step 1: Evaluating Construct Fidelity. Multiple-choice items are often criticized for their limited ability to produce authentic contexts associated with the constructs they target due to their inauthentic response format (Bryant, 2017; Gifford, 2017; Scully, 2017; Sireci & Zenisky, 2006). A TEI judged to have low construct fidelity indicates that "the context in which responses are produced and/or the method used to produce a response do not authentically reflect how the construct is typically applied outside of a testing situation" (Russell, 2016, p. 25). Consequently, a low fidelity TEI does not provide any gain (and perhaps even a loss) on construct representation compared to an MCI counterpart. Considering that improved construct representation is a cornerstone "promise" of TEIs, a lack of construct fidelity is unacceptable while moderate or high fidelity are desirable. For this reason, when a TEI has low fidelity the protocol recommends that one should immediately defer to the MC equivalent without considering any other indicators.

Step 2: Evaluating Difficulty. Evaluating difficulty requires consideration of both any change in difficulty and the role, if any, that construct representation seems to play in that change. A difference in difficulty between a stem-equivalent TEI-MCI pair is not inherently beneficial or detrimental. If a TEI is significantly more difficult than its stem-equivalent multiple-choice counterpart it may be due to the TEI assessing the targeted construct at a higher cognitive level (e.g., plotting a function instead of choosing one from a set of options —construction vs. recognition) or decreasing test-taker guessing. Similarly, a significantly easier TEI could be due to an authentic context that allows test-takers to demonstrate their knowledge more clearly or to a reduction in cognitive load (e.g., a classification drag-and-drop reduces reading load by removing repetition and

complex comparisons across MC options). In both cases, however, it is also possible that significant changes in difficulty are a result of unintended construct irrelevant variance. Thus, the desirability or undesirability of a significant change in difficulty is dependent on the context and purpose of the item.

If a significant difference in difficulty is observed, it is essential to consider whether this difference is construct-relevant (i.e., explained by the increase on construct fidelity) or not. The construct fidelity ratings evaluated in Step 1 provide useful context to make this judgment. If an item has been found to have high construct fidelity it is an indication that both the context and the interactions presented by the item are authentic and improve construct representation. Consequently, if high construct fidelity is observed, significant differences between difficulty parameters within an item pair are assumed to be construct-relevant (i.e., high fidelity cannot lead to a construct-irrelevant change in difficulty). In contrast, moderate fidelity indicates that while the context is authentic the interactions used to provide a response are not. Thus, one should evaluate whether the inauthentic interactions are associated with the difference in difficulty; if so, this difference is considered construct-irrelevant.

In this step, a construct-irrelevant significant difference in difficulty is deemed a detrimental outcome and the multiple-choice format is favored. Otherwise, if the change in difficulty is not significant or significant and deemed to be associated with increased construct fidelity (neutral and beneficial outcomes), then the protocol proceeds to consider item discrimination.

Step 3: Evaluating Discrimination. For nearly all types of tests, an increase in item discrimination is desirable. For this reason, any significant decrease in discrimination

produced by the TEI produces a recommendation to use the MC version. If the TEI significantly increases discrimination or if there is no impact on discrimination, one should then consider the impact that the use of the TEI has on information and efficiency.

Step 4: Evaluating Efficiency. Measurement efficiency values ($E(I_j)/\text{min}$) are also point estimates that are directly comparable. No statistical tests are available to evaluate the statistical significance of differences in this indicator; thus it is compared at face value. To aid this comparison, a ratio expressing the *relative measurement efficiency* of the TEI to the MCI is calculated as

$$\frac{E(I_{TEI})/\text{min}}{E(I_{MCI})/\text{min}} \tag{17}$$

and compared to two criteria informed by Jodoin's analyses: 2 and 0.5. Values larger than 2 indicate the TEI statistic was meaningfully larger and values less than 0.5 indicate the MCI statistic was meaningfully larger (note that TEI:MCI ratios smaller than 0.5 are analogous to MCI:TEI ratios larger than 2). A relative measurement efficiency value below 0.5 is undesirable because it indicates that the TEI takes more time to respond relative to its expected information compared to the MCI. Consequently the multiple-choice version is recommended in this case. In contrast, values or relative measurement efficiency larger than 2 are desirable. If the value of relative measurement efficiency falls between 0.5 and 2 the difference between the measurement efficiency of both formats is considered negligible (i.e., a neutral outcome).

In review, evaluating construct fidelity was chosen as the first step because a judgment of low fidelity indicates no gain on the primary "promise" of TEIs, namely,

better construct representation (Bryant, 2017; Sireci & Zenisky, 2006). Evaluating difficulty was selected as the second step because the gains in fidelity observed in the first step provide context to assess whether observed significant differences in difficulty parameters are construct relevant or construct irrelevant, the latter being unacceptable. The third step, evaluating discrimination, follows naturally because a large decrease in discrimination may threaten properties of the test such as total information and reliability. Finally, relative measurement efficiency across item formats is evaluated in the last step.

At each step in the comparative measurement value protocol there is an automatic recommendation in favor of the multiple-choice format if an undesirable outcome is observed: low construct fidelity, a construct-irrelevant significant difference in item difficulty, significantly lower discrimination, or a meaningfully low value of relative measurement efficiency. Otherwise, if desirable or neutral outcomes are observed an ordinal score of 1 or 0 is assigned respectively. These scores are used to provide a final CMV rating ranging between 0 and 4. These ratings provide an indication of the increase in comparative measurement value a TEI provides in relation to its MCI counterpart. Ratings larger than 0 correspond to an increase in comparative measurement value while a rating of 0 indicates there was no impact on CMV.

CMV ratings correspond to the number of indicators in which a TEI shows a beneficial outcome. Ratings larger than 0 result in a recommendation that favors the use of the TEI due to the increase in comparative measurement value and the absence of undesirable outcomes resulting from the use of the TEI format. In contrast, a CMV rating of 0 represents a TEI that has moderate fidelity coupled with a net neutral outcome on the remaining indicators. Given the lack of desirable changes in difficulty, discrimination,

and efficiency, the marginal gain in construct fidelity is not sufficient for justifying the use of TEI over an MC counterpart and thus the multiple-choice format is recommended.

Applying the CMV protocol

Table 5.1 presents a summary of the comparisons between TEIs and stem-equivalent MCIs used in this study, the resulting judgments of the application of the CMV protocol, and the recommendation for which item to use. Note that two items, RO3 and RO4, would not be included in an operational test due to their erratic psychometric behavior (extreme difficulties and negative discrimination).

Of the items to which the comparative measurement protocol was applied, seven technology-enhanced items were judged to have no impact on CMV, one judged to have increased CMV, and two judged to have decreased CMV. The significant difference in difficulty parameters between items CL4 and MCCL4 was deemed construct-irrelevant and thus CL4 received a judgment of decreased CMV.

All rank-ordering items were deemed to have no impact on or decreased comparative measurement value. Classification items showed a similar pattern with the exception of one item (CL7) which showed increased CMV. Based on these comparisons, use of a multiple-choice format is recommended for all rank-ordering items and use of only one classification TEI is recommended. These results are consistent with comments made by the panelists who rated construct fidelity and who observed that several of the constructs targeted by the TEIs would be better addressed through multiple-choice questions.

Table 5.1

Item response format recommendations for items in the data collection instrument based on the CMV protocol

Item	Construct fidelity	IRT		Measurement efficiency[a]	CMV rating	CMV judgment	Recommended format
		b	*a*				
				Classification			
CL3	Moderate	-	-	0.92	0	No impact	MCI
CL4	Moderate	+*	+	1.71		Decreased	MCI
CL5	Moderate	+	+	0.99	0	No impact	MCI
CL6	Moderate	+	+	1.26	0	No impact	MCI
CL7	Moderate	-	+	2.78†	1	Increased	MCI
CL8	Moderate	-	+	0.98	0	No impact	MCI
				Rank-ordering			
RO3	Moderate	+*	-	13.45†			
RO4	Moderate	-*	+	$< .001$‡			
RO5	Moderate	-	-	0.99	0	No impact	MCI
RO6	Moderate	-	-*	0.65		Decreased	MCI
RO7	Moderate	-	-	0.83	0	No impact	MCI
RO8	Moderate	-	-	0.59	0	No impact	MCI

\+ TEI statistic was higher than MCI. - TEI statistic was lower than MCI.

*Difference between TEI and MCI statistic was considered statistically significant.

[a] Value of relative measurement efficiency i.e., $E(I_{TEI})$/min:$E(I_{MCI})$/min.

† Ratio across formats (TEI:MCI) was meaningfully large (i.e., > 2).

‡ Ratio across formats (TEI:MCI) was meaningfully small (i.e., < 0.5).

Usage of the CMV protocol

To use the comparative measurement value protocol to determine the best format for an item nine steps are recommended:

1. Develop stem-equivalent pairs of technology-enhanced and multiple-choice items. Have content experts review the items to ensure that both items target the same construct despite the differences in response interactions.

2. Convene a panel of experts to rate the construct fidelity of technology-enhanced items considered.

3. Administer the items to a sample of test-takers and score responses.
4. Calibrate data using an item response model. This study employed a 2-parameter logistic model; however, other models may be considered.
5. Estimate item difficulty and discrimination, expected information and measurement efficiency (expected information / median item time).
6. Evaluate whether any item pairs should be discarded due to negative discrimination parameters or other relevant considerations.
7. Determine whether there were significant item parameter differences between stem-equivalent items. This may be accomplished through the "3-sigma IRT" method used in this study to identify item parameter drift or other approaches available in the literature.
8. Calculate the ratio of measurement efficiency across formats (i.e., relative measurement efficiency).
9. Compare item characteristics within item pairs following the CMV protocol:

    Step 1: Evaluate construct fidelity.

    Step 2: Evaluate difference in difficulty (and if significant, whether it is construct-relevant).

    Step 3: Evaluate difference in discrimination.

    Step 4: Evaluate relative measurement efficiency.

Appendix E presents a map of all possible paths that result from the CMV protocol as a practical alternative to the decision tree presented in Figure 5.1.

Implications

Over the last two decades the educational assessment industry has developed technology-enhanced items with the intention of improving measurement of targeted constructs. However, there is a lack of literature regarding what "better" means when comparing TEIs to traditional item formats. The body of literature that examines comparison of TEI and MC items has several shortcomings. Foremost among these shortcomings is the use of items that target different content and constructs when comparing items with different response formats. In addition, although a useful tool for evaluating select aspects of a TEI, the TEI Utility Framework relies exclusively on subjective judgment and focuses on the evaluation of the item prompt and interaction space (i.e., the design) absent any consideration of psychometric properties.

This dissertation addressed some of the methodological shortcomings of the relevant limited body of work by employing stem-equivalent pairs of TEIs and MCIs which provided content-equivalence across items with different response interactions. This study also estimated several psychometric properties of both the TEI and MC items in a pair. Building on these analyses, this dissertation proposes a decision protocol that extends the TEI Utility Framework by using quantitative characteristics to evaluate the potential gains in measurement value that might occur when a multiple-choice item is replaced by a technology-enhanced item.

Overall, this dissertation develops theory and proposes methodology that supplements the growing literature about comparisons between innovative item types and traditional item formats. Test developers and future literature examining differences

between technology-enhanced items and traditional item formats may benefit from the methodology outlined in this work to better understand how the psychometric properties of TEIs may be used to inform when it is most appropriate to use these item formats. The protocol proposed through this study provides a clear argument-based rationale to understand situations in which test developers benefit from replacing a multiple-choice item with a technology-enhanced item. This protocol lays out the trade-offs associated with any decision (either favoring a TEI or MCI format) and may be utilized to inform an operational assessment's validity argument.

Testing programs may apply the methodology described in this dissertation by embedding this research in a regular field-testing exercise so long as stem-equivalent item pairs are developed. Doing so would mitigate the cost of recruiting participants that would be needed otherwise to conduct an independent study.

The application of the CMV protocol to the items used in this study resulted in most of the cases in a recommendation favoring the multiple-choice format over the TEI version. Although informative, these results do not extend to drag-and-drop items used in all operational testing programs for two main reasons: (a) the diversity of digital delivery platforms available produce different drag-and-drop layouts than the ones employed in this study and (b) construct relevance of the response interaction of a TEI is a central aspect of the comparative measurement value protocol. Consequently, assessment programs should rely on independent studies to draw conclusions about the comparative measurement value of the drag-and-drop formats (and other TEI interactions) being used considering their purpose and content framework as well as the capabilities of their digital delivery platforms.

Limitations

This study was limited by the characteristics of the participant sample. The reliability of data obtained from participants recruited through Amazon's MTurk has been studied in the context of surveys for social science research. It is unclear, however, if the same results hold for cognitive instruments such as the one employed for this study. MTurk's workers are compensated for the tasks they complete. Consequently, workers are motivated to complete as many tasks as they can in as short a period of time possible. Because the data collection instrument used may require more effort (and time) to complete than an attitudinal survey, careless responses by participants trying to finish the instrument quickly may have occurred more frequently than during a typical test administration. Additionally, despite the extensive efforts conducted to ensure data quality (i.e., attention control items, review of open-ended questions, and time spent), the possibility remains that participant responses that appeared to be valid were not. Moreover, the motivation of MTurk workers to complete more tasks in a short period of time may have been heightened at the time of this research. Originally, this study was to be conducted in classrooms in the state of Massachusetts with an instrument targeting mathematics and science at the 8th-grade level. The COVID-19 pandemic forced local authorities to close schools and this study was redesigned. Thus, despite the known limitations associated with MTurk, it was deemed the best recourse to recruit adult participants to answer a cognitive instrument. Nonetheless, this study was conducted in the summer of 2020, when a large group of people lost or experienced reductions in their main sources of income due to the COVID-19 pandemic. As a result, people who regularly used Amazons MTurk as a source of supplemental income may have started relying on it as their main source of

income and in turn may have been motivated to complete assignments with even greater speed and less attention.

Some characteristics of the data collection instrument also are a source of limitations for this study. First, because items addressed three different subjects (statistics, science, and mathematics) the meaning of the total scores was unclear. This, in turn limits the value of the discrimination item statistic (particularly in the CTT paradigm). In other words the data collection instrument was not a test designed to measure a clearly defined domain. Second, the expected time required to answer the instrument (maximum 30 minutes) coupled with the use of a common block to calibrate the IRT model, limited the number of item pairs (12) that informed the development of the comparative measurement value protocol. This limitation was also increased by the fact that two item pairs (RO3-MCCRO3 and RO4-MCCRO4) showed erratic psychometric properties (extreme difficulty and negative discrimination).

This study was limited to only one class of TEI interactions (drag-and-drop) and only two types of this interaction were employed (classification and rank-ordering). These interactions do not encompass all possible drag-and-drop response spaces. For example rank-ordering items presented participants with objects in a vertical list to organize. However, in current operational assessments it is also common to see drag-and-drop ordering items present the objects horizontally. Consequently, the generalizability of the results of this study is limited, both to drag-and-drop items as a class and TEIs as a whole. Similarly, all TEI stem-equivalent counterparts were 6-option multiple-choice items which limit the generalizability to other traditional selected-response formats.

The lack of variability in construct fidelity ratings resulting from the work of the panel limited the capacity of this dissertation to fully explore research question 2. Panelists commented that they did not have enough familiarity with the constructs targeted by some of the items and how they were taught at their academic levels (in particular the middle school science items). Moreover, the consensus discussion among panelists revealed that panelists could have benefitted from better training on the use of the construct fidelity coding guide (Russell & Moncaleano, 2019). Two main elements could improve the training: (a) a greater focus on discussing what makes a TEI interaction authentic; and (b) further practice with items of the content domain the raters will interact with when making their ratings. In sum, the resulting lack of variability in construct fidelity ratings could be explained by these two limitations of the panel's work.

Directions for Future Work

The results of this study provide multiple avenues to explore further the comparative measurement value of technology-enhanced items. The comparative measurement value protocol may be expanded to consider other selected-response formats other than multiple-choice questions. This effort may involve exploring how other psychometric models may be used as part of the protocol, for example, the 3-parameter model for dichotomous responses or models for polytomous scoring. Moreover, the application of the comparative measurement value protocol in this dissertation assumed that all multiple-choice items used had low or moderate construct fidelity (i.e., an inauthentic response interaction). However, it would be informative to judge the authenticity of the contexts presented by the MC items. The construct fidelity coding guide (Russell & Moncaleano, 2019) may be modified and applied to multiple-choice items to evaluate their construct

fidelity. Then, the first step of the CMV protocol could be updated to judge the gain in construct fidelity rather than the fidelity rating of the technology-enhanced item itself.

The protocol itself may be refined by exploring other indicators that would provide more nuance to comparative measurement value judgments such as expected information and cost of development. Expected information was not included in the protocol because empirical evidence suggested it provided redundant information relative to measurement efficiency. However, the relationship between these two item characteristics warrants further investigation given the small sample of items studied. Moreover, the relationship between discrimination and expected information could also be examined. According to theory, an increase in discrimination will lead to an increase in information. However, given that expected information is calculated as a weighted average of information (using ability estimates as the weights) higher discrimination does not always produce higher expected information (see item CL8 in Table 4.21).

An important difference between multiple-choice and technology-enhanced items is the cost associated with their development. Generally, it is believed that multiple-choice items are less expensive to develop than TEIs (Bryant, 2017; Gifford, 2017). However, it is also often noted that the development cost of TEIs declines over time (i.e., through repeated assessment cycles). Thus it would be valuable to include the cost of development as an indicator in the CMV decision protocol. Cost of item development could provide more nuance to the CMV protocol ratings. For example, if the cost of developing a TEI is relatively higher than developing an MCI, a higher increase in CMV would be desirable (e.g., CMV ratings of 3 or 4). Similarly, if the cost differential is low then a lower increase in CMV would be acceptable (e.g., CMV ratings of 1 or 2). However, a significant increase

in cost coupled with a low increase in CMV might not be acceptable. Cost of item development was not included as an indicator in this work as this data is not publicly available and development cost may be dependent on specific needs and resources of different assessment programs.

Finally, as mentioned above, the generalizability of the results to drag-and-drop items as a class and TEIs as a whole is limited. The CMV protocol may be applied to more drag-and-drop item interactions so that broader conclusions can be made about this class of TEI interactions and inform the use of this format in operational assessments. Such effort would complement a recent wave of research studies focusing on the impact of drag-and-drop interactions on student performance and response times (Arslan, 2020; Ponce, Mayer & Loyola, 2020; Ponce, Mayer, Sitthiworachart & Lopez, 2020). Future work could also employ the protocol described here to evaluate the comparative measurement value of other TEI response interaction spaces.

Conclusion

This dissertation was motivated by a strong belief that technology-enhanced items require more scrutiny regarding the contexts in which they should be used. Consequently, this work explored and developed an approach for judging the comparative measurement value of technology-enhanced formats with respect to the multiple-choice format. This study relied on stem-equivalent pairs of drag-and-drop and multiple-choice items administered to a sample of participants and used a 2-parameter IRT model to calibrate the results. The most valuable quantitative and qualitative characteristics to judge comparative measurement value were then identified and used to inform the construction of a decision protocol.

The resulting protocol was applied to the items used in this study and revealed that all rank-ordering drag-and-drop items used showed decreased measurement value while classification items showed a mixture of desirable and undesirable CMV judgments. In turn, these results indicate that in a real-life scenario the multiple-choice format would be favored for most of these item pairs. Although the generalizability of these results to drag-and-drop items as a class is limited, they do suggest that drag-and-drop items should be used sparingly given their lack of value compared to MC items. The proposed protocol provides clear evidence-based rationales that will inform validity arguments of operational assessment programs and the usage of common technology-enhanced formats. Finally, the protocol may be refined in the future by adapting it to include other selected response formats and psychometric models and including additional indicators such as item development cost.

Glossary of Terms

Comparative measurement value: A judgment regarding the benefit of using one item format versus an alternate format with respect to increasing construct fidelity and improving psychometric characteristics.

Construct fidelity: One of the three facets of the TEI Utility Framework (Russell, 2016). Encompasses two main components: the authenticity of the context presented by an item relative to real-world contexts and the authenticity of the interaction used to produce a response reflect methods to apply the targeted construct outside of a testing situation.

Efficiency: Label used in this work to refer to the fourth and last step of the Comparative Measurement Value protocol which evaluates relative measurement efficiency (see definition below).

Expected information: Average of the information function of an item evaluated at test-takers' ability estimates: $E(I_j) = \frac{1}{N} \sum_{i=1}^{N} I_j(\hat{\theta}_i)$.

Measurement efficiency: The ratio of expected information of an item to the median time spent by test-takers in that item, i.e., $E(I_j)/\text{min}$ (Jodoin, 2003).

Relative efficiency: The definition of this term varies by author, two main definitions exist, one introduced by Lord (1980) and a second one introduced by Jodoin (2003). Both approaches to calculate relative efficiency are used in this work and are referred to, for clarity, as the *relative efficiency ratio* and *relative measurement efficiency* respectively (see definitions below).

Relative efficiency ratio: The ratio of two information curves. Introduced by Lord (1980) as relative efficiency. The result of this ratio is a second function which may be plotted and interpreted by comparing it to a reference line equal to 1. Used in the works of Crabtree (2016) and Gutierrez (2009).

Relative expected information: The ratio of the expected information for two items differing response format. In this work, it was calculated comparing technology-enhanced items to multiple-choice items as $\frac{E(I_{TEI})}{E(I_{MCI})}$.

Relative measurement efficiency: The ratio of measurement efficiency for two items differing response format. In this work, it was calculated comparing technology-enhanced items to multiple-choice items as $\frac{E(I_{TEI})/\min}{E(I_{MCI})/\min}$. Referred to as "relative efficiency" in Jodoin (2003).

Technology-enhanced item: An item that requires test-takers to "demonstrate knowledge, skills, and abilities using response interactions that provide methods for producing responses other than selecting from a set of options or entering alphanumeric content" (Russell, 2016, p. 20).

Usability: One of the three facets of the TEI Utility Framework (Russell, 2016). Russell defines it as "the intuitive functionality of an interactive space and the easy with which a novice user can produce and modify responses with minimal mouse or finger and/or response control selections" (Russell, 2016, p. 25).

References

Ackerman, T. A. & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement*, *12*(2), 117–128.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Allen, M. J. & Yen, W. M. (2002). *Introduction to measurement theory*. Waveland Press.

Al-Rukban, M. O. (2006). Guidelines for the construction of multiple-choice questions tests. *Journal of family and community medicine*, *13*(3), 125–133.

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

Anderson, L. W. (Ed.), Krathwohl, D. R. (Ed.), Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives*. Longman.

Arslan, B., Jiang, Y., Keehner, M., Gong, T., Katz, I. R., & Yan, F. (2020). The effect of drag-and-drop item features on test-taker performance and response strategies. *Educational Measurement: Issues and Practice*, *39*(2), 96–106.

Association of Test Publishers & Institute for Credentialing Excellence. (2017). *Innovative Item Types: A white paper & portfolio*. Association of Test Publishers.

Baker, F. B. (2001). *The Basics of Item Response Theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Barger, P., Behrend, T. S., Sharek, D. J., & Sinar, E. F. (2011). I-O and the crowd: Frequently asked questions about using Mechanical Turk for research. *The Industrial-Organizational Psychologist*, *49*(2), 12–17.

Barnett-Foster, D. & Nagy, P. (1996). Undergraduate student response strategies to test questions of varying format. *Higher Education*, *32*(2), 177–198.

Behrens, J. T., DiCerbo, K. E., & Foltz, P. W. (2019). Assessment of complex performances in digital environments. *The Annals of the American Academy*, *683*.

Benedetto, S., Carbone, A., Drai-Zerbib, V., Pedrotti, M., & Baccino, T. (2014). Effects of luminance and illuminance on visual fatigue and arousal during digital reading. *Computers in Human Behavior*, *41*, 112–119.

Bennett, R. E. (1993). On the meaning of constructed response. In W. C. Ward & R. E. Bennett (Eds.), *Construction versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment* (pp. 1–27). Lawrence Erlbaum Associates.

Bennett, R. E. (1998). *Reinventing assessment. speculations on the future of large-scale educational testing. a policy perspective*. Educational Testing Service.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, *39*(1), 370–407.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it Matter if I Take my Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP. *Journal of Technology, Learning, and Assessment*, *6*(9).

Bennett, R. E., Morley, M., & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized-adaptive tests. *Applied Psychological Measurement*, *24*, 294–309.

Bennett, R. E., Morley, M., Quardt, D., & Rock, D. A. (2000). Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning. *Applied Measurement in Education*, *13*, 303–322.

Bennett, R. E. & Rock, D. A. (1995). Generalizability, validity, and examinee perceptions of a computer-delivered formulating-hypotheses test. *Journal of Educational Measurement*, *32*, 19–36.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1990). The relationship of expert-system score constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement*, *14*, 151–162.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, *28*(1), 77–92.

Berg, C. A. & Smith, P. (1994). Assessing students' abilities to construct and interpret line graphs: Disparities between multiple-choice and free-response instruments. *Science Education*, *78*(6), 527–554.

Bergstrom, B. A. (1992, April 20-24). *Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis.* [Paper presentation]. American Educational Research Association Annual Meeting, San Francisco, CA, United States.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, *20*, 351–368.

Berk, R. A. (1980). *Criterion-referenced measurement: The state of the art.* The Johns Hopkins University Press.

Bernard, M., Chaparro, B., & Thomasson, R. (2000). Finding information on the web: Does the amount of whitespace really matter? *Usability News*, *2*(1).

Bernard, M., Lida, B., Riley, S., Hackler, T., & Janzen, K. (2002). A comparison of popular online fonts: Which size and type is best? *Usability News*, *4*(1).

Bernard, M. & Mills, M. (2000). So, What size and type of font should I use on my website? *Usability News*, *2*(2).

Bernard, M., Mills, M., Peterson, M., & Storrer, K. (2001). A comparison of popular online fonts: Which is best and when? *Usability News*, *3*(1).

Binet, A. & Simon, T. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L'annee Psychologique*, *12*, 191–244.

Birenbaum, M. & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats-It does make a difference for diagnostic purposes. *Applied Psychological Measurement*, *11*, 385–395.

Birenbaum, M., Tatsuoka, K. K., & Gutvirtz, Y. (1992). Effects of response format on diagnostic assessment of scholastic achievement. *Applied Psychological Measurement*, *16*(4), 353–363.

Blazer, C. (2010). Computer based assessments. *Information Capsule*, *918*.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational objectives. the classification of educational goals, Handbook 1: Cognitive domain.* Longmans.

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences.* Springer Science & Business Media.

Booth, J. F. (1998). The user interface in computer-based selection and assessment: Applied and theoretical problematics of an evolving technology. *International Journal of Selection and Assessment*, *6*, 61–81.

Borja, R. R. (2002, May 9). One state's digital quest. *Education Week*, *21*(35), 47–52.

Boyle, A. & Hutchison, D. (2009). Sophisticated Tasks in e-assessment: What are they and what are their benefits? *Assessment & Evaluation in Higher Education*, *34*(3), 305–319.

Bracht, G. H. & Hopkins, K. D. (1970). The communality of essay and objective tests of academic achievement. *Educational and Psychological Measurement*, *30*, 359–364.

Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Boner, M. W. (1994). Performance versus objective testing and gender: An exploratory study of an advanced placement History examination. *Journal of Educational Measurement*, *31*(4), 275–293.

Breland, H. M. & Gaynor, G. L. (1979). A comparison of direct and indirect assessments of writing skill. *Journal of Educational Measurement*, *16*(2), 119–128.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple- choice formats. *Journal of Educational Measurement*, *29*, 253–271.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, *16*, 191–205.

Bridgeman, B. & Lewis, C. (1994). The relationship of essay and multiple-choice scores with college courses. *Journal of Educational Measurement*, *31*, 37–50.

Bridgeman, B. & Rock, D. A. (1993). Relationship among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement*, *30*, 313–329.

Briel, J. & Michel, R. (2014). Revisiting the GRE general test. In C. Wendler & B. Bridgeman (Eds.), *The Research Foundation for the GRE revised General Test: A compendium of studies*. Educational Testing Service.

Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research & Evaluation*, *22*(1), 1–10.

Bugbee, A. C., Jr. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, *28*(3), 282.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). *The four generations of computerized educational measurement.* (ETS RR-88-35). Research Report Series. Educational Testing Service.

Cai, L., Du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: User guide.* Scientific Software International.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology*, *59*, 173–194.

Carcelli, L., Taylor, C., & White, K. (1980). *The effect of item format on phonics subtest scores of standardized reading achievement tests.* ERIC Document Reproduction Service No. ED206654.

Carson, J. (1993, June). Army alpha, army brass, and the search for army intelligence. *The History of Science Society*, *84*(2), 278–309.

Chang, P.-C., Chou, S.-Y., & Shieh, K.-K. (2013). Reading performance and visual fatigue when using electronic paper displays in long-duration reading tasks under various lighting conditions. *Displays*, *34*, 208–214.

Chaparro, B. S., Baker, J. R., Shaikh, A. D., Hull, S., & Brady, L. (2004). Reading Online Text: A comparison of four white space layouts. *Usability News*, *6*(2).

Chaparro, B. S., Phan, M., & Jardina, J. R. (2013, September 3 – October 4). *Usability and performance of tablet keyboards: Microsoft Surface vs. Apple iPad.* [Paper Presentation]. Human Factors and Ergonomics Society International Annual Meeting, San Diego, CA, United States.

Chaparro, B. S., Phan, M., Siu, C., & Jardina, J. R. (2014). User performance and satisfaction of tablet physical keyboards. *Journal of Usability Studies*, *9*(2), 70–80.

Chaparro, B. S., Shaikh, A. D., & Baker, J. R. (2005). Reading online text with a poor layout: Is performance worse? *Usability News*, *7*(1).

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), 213–255.

Chen, J. & Perie, M. (2018). Comparability within computer-based assessment: Does screen size matter? *Computers in the schools*, *35*(4), 268–283.

Chen, W. H. & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in Organizational Psychology: An evaluation and practical recommendation. *Journal of Business and Psychology*, *32*, 347–361.

Choi, S. W. & Tinkler, T. (2002, April 2-4). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K12 setting.* [Paper Presentation]. National Council on Measurement in Education Annual Meeting, New Orleans, LA, United States.

Cirn, J. T. (1986). True-false versus short answer questions. *College Teaching*, *34*, 34–37.

Clariana, R. & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, *33*(5), 593–602.

Clarke, M. M., Madaus, G. F., Horn, C. L., & Ramos, M. A. (2000). Retrospective on educational testing and assessment in the 20th century. *Journal of Curriculum Studies*, *32*(2), 159–181.

Coffman, W. E. (1966). On the validity of essay tests of achievement. *Journal of Educational Measurement*, *3*, 151–156.

College Board. (n.d.). *Sample questions.* Retrieved June 3, 2020, from https://collegereadiness.collegeboard.org/sample-questions

Computer Hope. (2019). *Pinch-to-zoom.* Retrieved January 4, 2020, from https://www.computerhope.com/jargon/p/pinch-to-zoom.htm

Copperud, C. (1979). *The test design handbook.* Educational Technology Publications.

Cortina, J. M. (1993). What is Coefficient Alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104.

Coulson, J. E. & Silberman, H. F. (1960). Effects of three variables in a teaching machine. *Journal of Educational Psychology*, *51*(5), 135–143.

Crabtree, A. R. (2016). *Psychometric properties of technology-enhanced item formats: An evaluation of construct validity and technical characteristics.* [Doctoral dissertation,

The University of Iowa] The University of Iowa's Institutional Repository. Retrieved from https://ir.uiowa.edu/cgi/viewcontent.cgi?article=6406&context=etd

Crocker, L. (1992). Item analysis. *Encyclopedia of educational research*, 652–657.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J. & Warrington, W. G. (1952). Efficiency of multiple-choice tests as a function of spread of item difficulties. *Psychometrika*, *17*(2), 127–147.

Davey, T., Godwin, J., & Mittelholtz, D. (1997). Developing and scoring an innovative computerized writing assessment. *Journal of Educational Measurement*, *34*, 21–42.

Davey, T. & Pitoniak, M. (2006). Designing computerized adaptive tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 543–574). Routledge.

Davis, F. B. & Fifer, G. (1959). The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, *14*(2), 159–170.

Davis, L. L., Kong, X., McBride, Y., & Morrison, K. M. (2017). Device Comparability of Tablets and Computers for assessment purposes. *Applied Measurement in Education*, *30*(1), 16–26.

Davis, L. L., Orr, A., Kong, X., & Lin, C. (2015). Assessing student writing on tablets. *Educational Assessment*, *20*, 180–198.

de Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

De Beer, M. (2004). Use of Differential Item Functioning (DIF) analysis for bias analysis in test construction. *SA Journal of Industrial Psychology*, *30*(4), 52–58.

DePascale, C., Dadey, N., & Lyons, S. (2016). *Score comparability across computerized assessment delivery devices: Defining comparability, reviewing the literature, and providing recommendations for states when submitting to Title 1 Peer Review*. Council of Chief State School Officers.

Diedenhofen, B. (2016). *R package "cocor"*. (Version 1.1-3). [Computer software].

Dillon, A., Richardson, J., & McKnight, C. (1990). The effect of display size and text splitting on reading lengthy text from screen. *Behaviour and Information Technology*, *9*(3), 251–217.

Dimock, P. H. & Cormier, P. (1991). The effects of format differences and computer experience on performance and anxiety on a computer-administered test. *Measurement and Evaluation in Counseling and Development*, *24*.

Dolan, R. P., Goodman, J., Strain-Seymour, E., Adams, J., & Sethuraman, S. (2011). *Cognitive lab evaluation of innovative items in mathematics and English/language arts assessment of elementary, middle, and high school students*. Pearson Education.

Donoghue, J. R. (1994). An empirical investigation of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, *31*, 295–311.

Dorans, N. J. & Holland, P. W. (1991). *DIF Detection and Description: Mantel-Haenszel and Standardization*. (ETS RR-92-10). Research Report Series. Educational Testing Service.

Douglass, H. R. (1926). *Modern methods in high-school teaching*. Riverside Press.

Drasgow, F. & Mattern, K. (2006). New tests and new items: Opportunities and issues. In D. Bartram & R. K. Hambelton (Eds.), *Computer-based testing and the internet: Issues and advances*. John Wiley & Sons, Ltd.

Duchnicky, R. L. & Kolers, P. A. (1983). Readability of text scrolled on a visual display terminals as a function of window size. *Human Factors*, *25*, 683–692.

Duke-Williams, E. & King, T. (2001, July 2-3). *Using computer-aided assessment to test higher level learning*. [Paper Presentation]. 5th International Computer Assisted Assessment (CAA) Conference, Loughborough, UK.

Dyson, M. C. (2004). How physical text layout affects reading from screen. *Behaviour and Information Technology*, *23*(6), 377–393.

Dyson, M. C. & Haselgrove, M. (2001). The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human–Computer Studies*, *54*, 585–612.

Dyson, M. C. & Kipping, G. J. (1998). The effects of line length and method of movement on patterns of reading from screen. *Visible Language*, *32*, 150–181.

Ebel, R. L. (1954). Procedures for the analysis of classroom tests. *Educational and Psychological Measurement*, *14*(2), 352–364.

Ebel, R. L. (1965). *Measuring educational achievement*. Prentice Hall, Inc.

Ebel, R. L. (1967). The relation of item discrimination to test reliability. *Journal of Educational Measurement*, *4*(3), 125–128. Retrieved from www.jstor.org/stable/1434085

Ebel, R. L. (1972). *Essentials of educational measurement*. Prentice-Hall.

Ebel, R. L. & Frisbie, D. (1991). *Essentials of educational measurement*. Prentice-Hall.

Eberhart, T. (2015). *A comparison of Multiple-choice and Technology-enhanced item types administered on Computer versus iPad*. [Doctoral dissertation, University of Kansas]. KU ScholarWorks. Retrieved from https://kuscholarworks.ku.edu/bitstream/handle/1808/21674/Eberhart_ku_0099D_14325_DATA_1.pdf?sequence=1&isAllowed=y

Educational Testing Service. (2009). *Graduate record examinations: Practice general test # 1*. Educational Testing Service. Retrieved May 16, 2020, from https://www.ets.org/s/gre/accessible/GRE_Practice_Test_1_Quant_18_point.pdf

Educational Testing Service. (2016). *The Road Ahead for State Assessments: What the assessment consortia built, why it matters, and emerging options*. Retrieved April 10, 2020, from https://www.ets.org/s/k12/pdf/coming_together_the_road_ahead.pdf

Educational Testing Service. (2017). *Practice book for the paper-delivered GRE General Test* (2nd ed.). Educational Testing Service. Retrieved May 16, 2020, from https://www.ets.org/s/gre/pdf/practice_book_GRE_pb_revised_general_test.pdf

Edwards, V. B., Chronister, G., Bushweller, K., Skinner, R. A., & Bowman, D. H. (2003, May 8). Tech's answer to testing. *Education Week*, *22*(35), 8–10.

Enders, C. K. (2010). *Applied missing data analysis*. The Guildford Press.

Enright, M., Rock, D. A., & Bennett, R. E. (1998). Improving measurement for graduate admissions. *Journal of Educational Measurement*, *35*, 250–267.

Eurich, A. C. (1931). Four types of examinations compared and evaluated. *Journal of Educational Psychology*, *26*, 268–278.

Fabrigar, L. R. & Wegener, D. T. (2011). *Exploratory factor analysis*. Oxford University Press.

Famularo, L. (2007). (UMI No. 3283877)[Doctoral Dissertation, Boston College]. ProQuest Dissertations and Theses Global.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357–381.

Fishbein, B. (2018). *Preserving 20 years of TIMSS trend measurements: Early stages in the transition to the eTIMSS Assessment*. [Doctoral dissertation, Boston College]. BC eScholarship. Retrieved from https://dlib.bc.edu/islandora/object/bc-ir%3A107927

Fishbein, B., Foy, P., & Tyack, L. (2020). Reviewing the TIMSS 2019 achievement item statistics. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 10.1–10.70).

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 Item Equivalence Study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education*, *6*(11).

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

Fraser, C. & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Researcch*, *23*, 267–269. Retrieved from https://doi.org/10.1207/s15327906mbr2302_9

Fraser, C. & McDonald, R. P. (2012). *NOHARM 4 Manual*.

French, J. L. & Hale, R. L. (1990). A history of the development of psychological and educational testing. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children* (pp. 2–28). Guilford.

Frisbie, D. A. & Cantor, N. K. (1995). The validity of scores from alternative methods of assessing spelling achievement. *Journal of Educational Measurement*, *32*(1), 55–78.

Gaertner, M. N. & Briggs, D. C. (2009). *Detecting and addressing item parameter drift in IRT test equating contexts*. Center for Assessment.

Gallagher, C. J. (2003, March). Reconciling a tradition of testing with a new learning paradigm. *Educational Psychology Review*, *15*(1), 83–99.

Gay, L. (1980). The comparative effects of multiple-choice versus open-ended on retention. *Journal of Educational Measurement*, *17*, 45–50.

Gierl, M. J., Lai, H., Pugh, D., Touchie, C., Boulais, A.-P., & De Champlain, A. (2016). Evaluating the psychometrics characteristics of generated multiple-choice test items. *Applied Measurement in Education*, *29*(3), 196–210.

Gifford, C. (2017). *Technology-enhanced items in assessment*. A Pass Educational Group.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability*. New York College Entrance Examination Board.

Goldberg, A. L. & Pedulla, J. J. (2002). Performance differences according to test mode and computer familiarity on a practice graduate record exam. *Educational and Psychological Measurement*, *62*(6), 1053–1067.

Gould, S. (1981). *The mismeasure of man*. WW Norton & Company.

Government of Alberta. (n.d.-a). *Student Learning Assessment - Practice Questions*. Retrieved October 22, 2017, from https://public.education.alberta.ca/assessment/Home/Practice/175?LanguageCode=en-CA&GradeCode=03

Government of Alberta. (n.d.-b). *Student Learning Assessment - Released Questions*. Retrieved October 22, 2017, from https://public.education.alberta.ca/assessment/Home/Index/?LanguageCode=en-CA

Gutierrez, S. L. (2009). *Examining the psychometric properties of a multimedia innovative item format: Comparison of innovative and non-innovative versions of a situational judgment test*. [Unpublished Doctoral Dissertation]. James Madison University.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Allyn & Bacon.

Haladyna, T. M. (1999). *Developing and Validating Multiple-choice Test Items* (2nd ed.). Lawrence Erlbaum Associates Inc.

Haladyna, T. M. (2004). *Developing and Validating Multiple-choice Test Items* (3rd ed.). Psychology Press.

Haladyna, T. M. & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Halpin, G., Halpin, G., & Harrington, J. (1981). Retention in an actual classroom setting as a function of type and complexity of tests. *Educational Research Quarterly*, *6*(2), 45–52.

Hancock, G. R. (1992). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *Journal of Experimental Education*, *62*(2), 143–157.

Harke, D. J., Herron, J. D., & Leftler, R. W. (1972). Comparison of a randomized multiple- choice format with a written on-hour physics problem test. *Science Education*, *56*, 563–565.

Harmes, J. C. & Wise, S. L. (2016). Assessing engagement during the online assessment of real-world skills. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 805–824). IGI Global.

Hartley, J. (1987). Designing electronic text: The role of print-based research. *Educational Communication and Technology*, *35*(1), 3–17.

Hase, H. D. & Goldberg, L. R. (1967). Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin*, *67*(4), 231–248.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied psychological measurement*, *9*(2), 139–164.

Hawkes, H. E., Lindquist, E. F., & Mann, C. R. (Eds). (1936). *The construction and use of achievement examinations: A manual for secondary school teachers*. Houghton Mifflin Company.

Heim, A. W. & Watts, K. P. (1967). An experiment on multiple-choice versus open-ended answering in a vocabulary test. *British Journal of Educational Psychology*, *37*, 339–346.

Henrysson, S. (1962). The relation between factor loadings and biserial correlations in item analysis. *Psychometrika*, *27*, 419–424.

Hetter, R. D., Segall, D. O., & Bloxom, B. M. (1997). Evaluating item calibration medium in computerized adaptive testing. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 161–167). American Psychological Association.

Hills, J. R. (1976). *Measurement and evaluation in the classroom*. Merrill Publishing Company.

Ho, S. Y., Lowrie, T., & Logan, T. (2015). *NAPLAN Online 2014 Development Study Cognitive Interviews Research Activity 3: Technically Enhanced Items (Numeracy)*. University of Canberra, Australia, prepared for the Australian Curriculum, Assessment and Reporting Authority. Retrieved from https://www.nap.edu.au/_resources/2014_NASOP_Technically_Enhanced_Items_Research_Report_UC_Numeracy.pdf

Hogan, T. P. (1981). *Relationship between free-response and choice-type tests of achievement: A review of the literature*. National Assessment of Educational Progress. (ERIC Document Reproduction Service No. ED 224 811).

Hogan, T. P. & Mishler, C. (1980). Relationships between essay tests and objective tests of language skills for elementary school students. *Journal of Educational Measurement*, *17*(3), 219–227.

Hopkins, C. D. & Antes, R. L. (1979). *Classroom testing construction*. F. E. Peacock Publishers, Inc.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, *5*(2).

Horn, J. L. (1966). Some characteristics of classroom examinations. *Journal of Educational Measurement*, *3*, 292–295.

Huff, K. L. & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice*, *20*, 16–25.

Hurd, A. W. (1932). Comparison of short answer and multiple-choice tests covering identical subject content. *Journal of Educational Research*, *26*, 28–30.

Hurlbut, D. (1954). The relative value of recall and recognition techniques for measuring precise knowledge of word meaning-nouns, verbs, adjectives. *Journal of Educational Research*, *47*, 561–576.

IMS Global Learning Consortium. (2020). *IMS Question and Test Interoperability (QTI): Assessment, Section and Item Information Model v3*. Retrieved from http://www.imsglobal.org/sites/default/files/spec/qti/v3/info/index.html#PCInt

Jewsbury, P., Finnegan, R., Xi, N., Jia, Y., Rust, K., & Burg, S. (2020). *2017 NAEP transition to digitally based assessments in mathematics and reading at grades 4 and 8: Mode evaluation study*. National Assessment of Educational Progress. Retrieved from https://nces.ed.gov/nationsreportcard/subject/publications/main2020/pdf/transitional_whitepaper.pdf

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, *40*(1), 1–15. Retrieved from https://www.jstor.org/stable/1435051

Jodoin, M. G. & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure with Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, *14*(4), 329–349.

Kamenetz, A. (2015). *The test*. Public Affairs.

Kelly, F. J. (1915). The Kansas silent reading test. *Studies by the Bureau of Educational Measurement and Standards*, *3*, 1–38.

Kinder, J. S. (1925, May). Supplementing our examinations. *Education*, *45*, 557–566.

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education*, *22*(1), 22–37.

Kinney, L. B. & Eurich, A. C. (1932). A summary of investigations comparing different types of tests. *School and Society*, *36*, 540–544.

Klein, A. (2015, April 10). No child left behind: An overview. *Education Week*. Retrieved from https://www.edweek.org/ew/section/multimedia/no-child-left-behind-overview-definition-summary.html

Koch, D. A. (1993). Testing goes graphical. *Journal of Interactive Instruction Development*, *5*, 14–21.

Kolers, P., Duchnicky, R. L., & Ferguson, D. C. (1981). Eye movement of readability of CRT displays. *Human Factors*, *23*, 517–527.

Kruk, R. S. & Muter, P. (1984). Reading of continuous text on video screens. *Human Factors*, *26*, 339–345. Retrieved December 17, 2019, from http://www2.psych.utoronto.ca/users/muter/Abs1984b.htm

Lay, Y.-K., Ko, Y.-H., Shieh, K.-K., Lee, D.-S., Yeh, Y.-Y., & Yang, T.-C. (2012). Visual performance and visual fatigue of long period reading on electronic paper displays. *Journal of Ergonomic Study*, *14*(2), 119–126.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing*, *6*(1), 1–24.

Lentz, T. F., Hirshstein, B., & Finch, F. H. (1932). Evaluation of methods of evaluating test items. *Journal of Educational Psychology*, *23*(5), 344.

Lindeman, R. H. & Merenda, P. F. (1979). *Educational measurement*. Pearson Scott Foresman.

Ling, G. (2016). Does it matter whether one takes a test on an iPad or a desktop computer. *International Journal of Testing*, *16*, 352–377.

Ling, G. & Bridgeman, B. (2013). Writing essays on a laptop or a desktop computer: Does it matter? *Journal of International Testing*, *13*(2), 105–122.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198–1202.

Livingston, S. A. (2006). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 421–441). Routledge.

Lord, F. M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, *17*(2), 181–194.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Erlbaum Associates.

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

Lovelace, E. A. & Southall, S. D. (1983). Memory for words in prose and their locations on the page. *Memory and Cognition*, *11*, 429–434.

Loyd, B. H. & Steele, J. (1986). Assessment of reading comprehension: A comparison of constructs. *Reading Psychology*, *7*, 1–10.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed-response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, *30*, 234–250.

Madaus, G. F. & O'Dwyer, L. M. (1999). A short history of performance assessment. *Phi Delta Kappa*, *80*(9), 688–695.

Madaus, G. F., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing.* Information Age Publishing, Inc.

Magill, W. H. (1934). The influence of the form of item on the validity of achievement tests. *Journal of Educational Psychology*, *25*, 21–28.

Margolin, S. J., Driscoll, C., Toland, M. J., & Kegler, J. L. (2013). E-readers, computer screens, or paper: Does reading comprehension change across media platforms? *Applied Cognitive Psychology*, *27*(4), 512–519.

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. 79–91). TIMSS & PIRLS International Study Center. Retrieved from http://timssandpirls.bc.edu/timss2019/frameworks/

Martinez, M. E. (1991). A comparison of multiple-choice and constructed response figural response items. *Journal of Educational Measurement*, *28*, 131–145.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34* (4), 207–218.

Massachusetts Department of Elementary and Secondary Education. (n.d.-a). *Released items*. MCAS resource center. Retrieved August 9, 2020, from http://mcas.pearsonsupport.com/released-items/

Massachusetts Department of Elementary and Secondary Education. (n.d.-b). *Student tutorial, practice tests and other resources*. MCAS resource center. Retrieved August 9, 2020, from http://mcas.pearsonsupport.com/student/

Maydeu-Olivares, A. & Joe, H. (2005). Limited and full information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009–1020.

Maydeu-Olivares, A. & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713–732.

Mazzeo, J. & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests*. (Report No. 88-8). College Entrance Examination Board.

McBride, J. R. (1979). *Adaptive mental testing: The state of the art*. (No. ARI-TR-423). Army Research Institute for the Behavioral and Social Sciences.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambelton (Eds.), *Handbook of modern item response theory* (pp. 257–269). Springer.

McDonald, R. P. (1999). *Test Theory: A unified treatment*. Erlbaum.

McMullin, J., Varnhagen, C. K., Heng, P., & Apedoe, X. (2002). Effects of surrounding information and line length on text comprehension from the web. *Canadian Journal of Learning and Technology*, *28*, 19–29.

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*, 449–458.

Measured Progress. (2015). *Technology-enhanced items*. Retrieved February 17, 2020, from https://www.measuredprogress.org/wp-content/uploads/2015/08/TEI-Flyer.pdf

Measured Progress & Educational Testing Service. (2012). *Smarter balanced assessment consortium: Technology enhanced items guidelines*.

Millman, J. & Setijadi. (1966). A comparison of performance of American and Indonesian students on three types of test items. *The Journal of Educational Research*, *59*(6), 273–275.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design*. (ETS RR-03-16). Research Report Series. Educational Testing Service.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds Ratio, Delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics*, *32*(1), 92–109.

Monahan, T. (1998). The Rise of Standardized Educational Testing in the U.S.: A Bibliographic Overview. *Rensselaer Polytechnic Institute Department of Science and Technology Studies*. Retrieved from https://kefk.pw/the_rise_of_standardized_educational_testing_in_the_us_a_bibliographic.pdf

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.

Mosier, C. I. (1936). A note on item analysis and the criterion of internal consistency. *Psychometrika*, *1*(4), 275–282.

Moss, P. A., Cole, N. S., & Khampalikit, C. (1982). A comparison of procedures to assess written language skills at Grades 4, 7, and 10. *Journal of Educational Measurement*, *19*(1), 37–47.

National Assessment of Educational Progress. (2018). *Digitally based assessments*. Retrieved from https://nces.ed.gov/nationsreportcard/dba/

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards: Mathematics*. National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C.

Neill, J. A. & Jackson, D. N. (1970). An evaluation of item selection strategies in personality scale construction. *Educational and Psychological Measurement*, *30*, 647–661.

Nielsen, J. (2012, January 3). *Usability 101: Introduction to usability*. Nielsen Norman Group. Retrieved from http://www.useit.com/alertbox/20030825.html

Nitko, A. J. & Hsu, T. (1984). *Item analysis appropriate for domain-referenced classroom testing*. University of Pittsburgh.

Nydick, S. W. (2015). *R package "catIrt"*. (Version 0.5-0). [Computer software].

Odell, C. (1928). *Traditional examinations and new-type tests*. The Century Co.

O'Leary, M., Scully, D., Karakolidis, A., & Pitsia, V. (2018). The state-of-the-art in digital technology-based assessment. *European Journal of Education*, *53*, 1–16.

Oosterhof, A. C. & Coats, P. K. (1984). Comparison of difficulties and reliabilities of quantitative word problems in completion and multiple-choice item formats. *Applied Psychological Measurement*, *8*, 287–294.

Organization for Economic Co-operation and Development. (2010). *PISA Computer-Based Assessment of Student Skills on Science*. Retrieved from https://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/pisacomputer-basedassessmentofstudentskillsinscience.htm

Orlando, M. & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64.

Orlando, M. & Thissen, D. (2003). Further investigation of the performance of S-X^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289–298.

Paolacci, G. & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, *23*(3), 184–188.

Parshal, C. G. & Harmes, J. C. (2007, June 7-8). *Designing templates based on a taxonomy of innovative items*. [Keynote Address]. GMAC Conference on Computer Adaptive Testing. Minneapolis, MN, United States.

Parshall, C. G. (2002). Item development and pretesting in a CBT environment. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 119–142). Laurence Erlbaum Associates, Publishers.

Parshall, C. G. & Becker, K. A. (2008, July 14-16). *Beyond the technology: Developing innovative items.* [Paper Presentation]. Bi-annual meeting of the International Test Commission, Liverpool, UK.

Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative item types for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 129–148). Kluwer Academic Publishers.

Parshall, C. G. & Guille, R. A. (2016). Managing ongoing changes to the test: Agile strategies for continuous innovation. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 1–22). Routledge.

Parshall, C. G. & Harmes, J. C. (2014). Improving the quality of innovative item types: Four tasks for design and development. *Journal of Applied Testing Technology*, *10*(1), 1–20.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. (2010). Innovative items for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 215–230). Kluwer Academic Publishers.

Parshall, C. G., Stewart, R., & Ritter, J. (1996). *Innovations: Sound, graphics, and alternative response modes.* [Paper Presentation]. National Council on Measurement in Education Annual Meeting, New York, New York, United States.

Partnership for Assessment of Readiness for College and Careers. (n.d.). *Practice tests.* Retrieved January 24, 2018, from https://parcc.pearson.com/practice-tests/

Paterson, D. G. (1926). Do new and old type examinations measure different mental functions? *School and Society*, *24*(608), 246–248.

Poggio, J. & McJunkin, L. (2012). History, Current Practice, Perspectives and what the Future Holds for Computer Based Assessment in K-12 Education. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessments: Recent history and predictions for the future* (pp. 25–53). Information Age Publishing.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment*, *2*(6).

Ponce, H. R., Mayer, R. E., & Loyola, M. S. (2020). Effects on test performance and efficiency of technology-enhanced items: An analysis of drag-and-drop response interactions. *Journal of Educational Computing Research*, *0*(0), 1–27.

Ponce, H. R., Mayer, R. E., Sitthiworachart, J., & Lopez, M. J. (2020). Effects on response time and accuracy of technology-enhanced cloze tests: An eye-tracking study. *Educational Technology Research and Development*, *68*(5), 2033–2053.

Powers, D. E. & Potenza, M. T. (1996). *Comparability of testing using laptop and desktop computers*. (ETS Report No. RR-96-15). Educational Testing Service.

Privitera, G. J. (2017). *Statistics for the behavioral sciences* (3rd ed.). Sage.

Qian, H., Woo, A., & Kim, D. (2017). Exploring the psychometric properties of innovative items in computerized adaptive testing. In H. Jiao & R. W. Lissitz (Eds.), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective* (pp. 97–118). Information Age Publishing.

Quellmalz, E. S., Capell, F. J., & Chou, C.-P. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement*, *19*(4), 241–258.

Race to the Top Fund Assessment Program, 75 C.F.R. §18171. (2010).

Rasch, G. (1966). An individualistic approach to item analysis. In P. Lazarsfeld & N. V. Henry (Eds.), *Readings in mathematical social science* (pp. 89–107). Science Research Association.

Rasch, G. (n.d.). *Probabilistic models for some intelligence and attainment*. Danish Institute for Educational Research.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14*, 127–137.

Richardson, M. W. (1936a). Notes on the rationale of item analysis. *Psychometrika*, *1*(1), 69–76.

Richardson, M. W. (1936b). The relation between the difficulty and the differential validity of a test. *Psychometrika*, *1*(2), 33–49.

Rizopoulos, D. (2018). *R package "ltm"*. (Version 1.1-1). [Computer software].

Robitzsch, A. (2020). *R package "sirt"*. (Version 3.9-4). [Computer software].

Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213–231). Routledge.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163–184.

Rouse, S. V. (2014). A reliability analysis of Mechanical Turk data. *Computers in Human Behavior*, *43*, 304–307.

Rowley, G. L. (1974). Which examinees are most favoured by the use of multiple-choice tests? *Journal of Educational Measurement*, *2*(1), 15–23.

Ruch, G. M. & Charles, J. W. (1928). A comparison of five types of objective tests in elementary psychology. *Journal of Applied Psychology*, *12*(4), 398–403.

Ruch, G. M. & Stoddard, G. D. (1925). Comparative reliabilities of five types of objective examinations. *Journal of Educational Psychology*, *16*(2), 89–103.

Russell, M. (2006). *Technology and Assessment: The tale of two interpretations.* (W. Heinecke, Ed.). Information Age Publishing Inc.

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

Russell, M., Mattson, D., Higgins, J., Hoffmann, T., Bebell, D., & Alcaya, C. (2011). *A primer to the accessible portable item profile (APIP) standards.* Minnesota Department of Education.

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. Retrieved from https://doi.org/10.1080/10627197.2019.1670055

Sahin, A. & Anil, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory and Practice*, *17*(1), 321–335.

Samejima, F. (1975, June 12-13). *Graded response model of the latent trait theory and tailored testing.* [Paper Presentation]. First conference on computerized adaptive testing, Washington, D. C., United States.

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project, research and development series.* (NCES 2005–457). U.S. Department of Education, National Center for Education Statistics. U.S. Government Printing Office.

Santos, P., Hernandez-Leo, D., Perez-Sanagustin, M., & Blat, J. (2012). Modeling the computing based testing domain extending IMS QTI: Framework models and exemplary implementations. *Computers in Human Behavior*, *28*, 1648–1662.

Sax, G. & Collet, L. S. (1968). An empirical comparison of the effects of recall and multiple-choice tests on student achievement. *Journal of Educational Measurement*, *5*(2), 169–173.

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), 1–43. Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1653

Schmeiser, C. B. & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Praeger Publishers.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research & Evaluation*, *22*(4).

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* (pp. 329–348). Routledge.

Smarter Balanced. (n.d.). *Practice tests*. Retrieved October 24, 2017, from https://practice.smarterbalanced.org/student/

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*, 66–78.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989, Fall). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement*, *26*(3), 261–271.

Strain-Seymour, E., Way, W. D., & Dolan, R. P. (2009). *Strategies and processes for developing innovative items in large-scale assessments*. Research Report. Pearson Education.

Swanson, C. B. (2006, May 9). Tracking U. S. Trends. *Education Week*, *25*(35), 50–53.

Terman, L. M. (1916). *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet-Simon*

intelligence scale. Houghton Mifflin. Retrieved from https://doi.org/10.1037/10014-000

Thissen, D. (1976). Information in wrong responses to Raven progressive matrices. *Journal of Educational Measurement*, *13*, 201–214.

Thissen, D., Wainer, H., & Wang, X.-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, *31*(2), 113–123.

Thorndike, E. L., Bregman, M. V., Cobb, M. V., & Woodyard, E. (1927). *The measurement of intelligence*. Teachers College, Columbia University.

Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). *Computer-based testing: Practices and considerations*. (Synthesis Report 78). University of Minnesota National Center on Educational Outcomes.

Thurstone, T. G. (1932). The difficulty of a test and its diagnostic value. *Journal of Educational Psychology*, *23*(5), 335.

Tiegs, E. W. (1939). *Tests and measurement in the improvement of learning*. Houghton Mifflin Company.

TIMSS and PIRLS International Study Center. (2019). *About TIMSS 2019*. Retrieved April 15, 2020, from https://timssandpirls.bc.edu/timss2019/

Traub, R. E. & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement*, *1*, 355–369.

Travers, R. M. (1950). *How to make achievement tests*. The Odyssey Press.

Tullis, T. S., Boynton, J. L., & Hersh, H. (1995, May 7-11). *Readability of fonts in the Windows environment*. [Paper Presentation]. Conference on Human Factors in Computing. Chicago, IL, United States.

U. S. Department of Education. (2010). *Race to the top assessment program executive summary*. Washington, D.C.

Urry, V. W. (1974). *Computer-assisted testing: The calibration and evaluation of the verbal ability bank*. (Technical study 74-3). Research Section, Personnel Research and Development Center, U. S. Civil Service Commission.

van den Bergh, H. (1990). On the construct validity of multiple-choice items for reading comprehension. *Applied Psychological Measurement*, *14*, 1–12.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204.

Vector Psychometric Group, LLC. (2020). *IRTPRO*. (Version 5.20). [Computer Software]. Retrieved from https://vpgcentral.com/software/irtpro/

Vernon, P. E. (1962). The determinants of reading comprehension. *Educational and Psychological Measurement*, *9*, 430–449.

Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*, 103–118.

Wan, L. & Henly, G. A. (2012). Measurement properties of two innovative item formats in a computer based test. *Applied Measurement in Education*, *25*(1), 58–78.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, *67*(2), 219–238.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement*, *6*, 1–11.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. *Journal of Educational Measurement*, *17*(1), 11–29.

Way, W. D., Davis, L. L., Keng, L., & Strain-Seymour, E. (2016). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 260–284). Routledge.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states*. (Research Monograph No. 18). Council of Chief State School Officers.

Weeks, J. P., von Davier, M., & Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychological Test and Assessment Modeling*, *58*(4), 671–701.

Weitzman, E. & McNamara, W. J. (1949). *Constructing Classroom Examinations: A guide for teachers*. Stratford Press.

Wendler, C. L. & Walker, M. E. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 445–467). Routledge.

Wendt, A. & Harmes, J. C. (2009). Evaluating innovative items for NCLEX, Part I: Usability and pilot testing. *Nurse Educator*, *34*(2), 56–59.

Wilmut, J. (1975). Objective test analysis: Some criteria for item selection. *Research in Education*, *13*(1), 27–56.

Wilson, M. & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement*, *19*(1), 51–71.

Wise, S. L. & Kong, X. L. (2009). Response time effort: A new measure of student motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

Yerkes, R. M. (1921). *Psychological Examining in the United States Army*. American Psychological Association.

Zenderland, L. (1998). *Measuring Minds: Henry Herbert Goddard and the Origins of American Intelligence Testing*. Cambridge University Press.

Zenisky, A. L. & Sireci, S. G. (2013, April 28-30). *Innovative items to measure higher-order thinking: Development and validity considerations*. [Paper Presentation]. National Council on Measurement in Education Annual Meeting, San Francisco, CA, United States.

Ziefle, M. (1998). Effects of display resolution on visual performance. *Human Factors*, *40*(4), 554–568.

Zieky, M. (2003). *A DIF primer*. Educational Testing Service. Retrieved from https://www.ets.org/s/praxis/pdf/dif_primer.pdf

Zubin, J. (1934). The method of internal consistency for selecting test items. *Journal of Educational Psychology*, *25*(5), 345–356.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement*. (ETS RR-12-08). Research Report Series. Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RR-12-08.pdf

Appendix A - Common Traditional and Technology-Enhanced Item Formats

This appendix is divided in three sections. In the first section, common traditional item formats are described, including: (a) multiple-choice, (b) alternate option, (c) matching, (d) rank-ordering, (e) fill-in-the-blank, (f) location/identification, (g) detection and correction of errors, (h) problem or short answer, and (i) extended response. Afterwards common technology-enhanced item formats are described, including: (a) drag-and-drop, (b) plot points, (c) select text, (d) create frequency plots, (e) shade area, (f) create partition, (g) hot spot, (h) matching, and (i) in-line choice. This appendix concludes with figures showcasing the innovative items discussed in Jodoin's (2003) study.

Traditional Item formats

The following section describes the most common traditional item formats that have been prevalent in both large-scale and classroom tests throughout the 20th century. Item formats that can be objectively scored are presented first (both selected response and constructed response) followed by subjectively scored item formats. Examples of the earliest uses of these item formats may be found in "Traditional Examinations and New-type Tests" (Odell, 1928) and "Tests and Measurement in the Improvement of Learning" (Tiegs, 1939).

Multiple-Choice

A multiple-choice (MC) item is a prompt (or stem) followed by several possible response options, one of which is correct and the others are not (usually referred to as distractors). The stem can be worded as a direct question or an incomplete sentence that is completed by the option selected by the test-taker. The quintessential multiple-choice item presents

four options to the test-taker, but it is also common to present three options or more than four (two options are considered a different category of item format). Variations of the MC item include situations in which all the options are correct responses to some extent but one option is clearly better (best-answer format), items that allow test-takers to select more than one option as a response (multiple-response format), or multiple-choice items organized into sets that use one list of options for all items in the set; these item sets include a theme, an option list, a lead-in statement and multiple item stems (extended multiple-choice format; Al-Rukban, 2006).

Alternate Option

Also called the 2-option MC item format, this item format presents the test-taker with a stem that is answered with one of two possible options. This item format was introduced as the True/False item format, where a test-taker evaluated statements on whether they were true or false. A variation of the True/False format added a third option, "Neither," for when the statement was neither true or false. A further variation added a fourth option to capture statements that required more information to be evaluated accurately. This item format was later relabeled the "alternate option" item format when further variations included other sets of options beyond True/False, such as Yes/No and Correct/Incorrect. Like with MC items, the extended alternate option item format became common, where multiple of statements were presented simultaneously requiring the tests-taker to judge all of them under the same criteria (e.g., true/false).

Matching

Matching items present two lists to the test-takers and require them to match elements of one list to the other according to some criterion or instruction explained in the item

prompt. A matching item typically includes two lists, one of which is numbered and the other which is preceded by a blank space in which the examinee records a response. In each of these spaces, test-takers write down the numeral that corresponds to the element of the numerated list they wish to match. Originally, matching items presented lists of equal length where each element of a list had a pair, but variations introduced lists of unequal-length with the intention of reducing the ability of test-takers to determine the response to the final item using the process of elimination. Further variations of this item format included items with more than two lists and allowing test-takers to respond by connecting elements between lists with lines.

Rank-ordering

The rank order item format asks the test-taker to order several statements. The ordering of the statements can follow any criteria such as chronological or logical order. These items present a blank space either at the beginning or the end of each statement where test-takers are expected to write numerals to indicate their ordering of the statements. Variations of this item format may replace statements for other stimuli, such as diagrams or drawings.

Fill-in-the-Blank

This item format presents students with a sentence or a paragraph with one or more embedded blank spaces corresponding to a single word each. Test-takers are required to fill in these blanks with appropriate words. This item format can be constrained by offering test-takers options from which to select to complete the text. For example, some tests present two words that could fill in the blank in a parenthesis following the blank (e.g., two conjugations of the same verb). Other tests provide something closer to a

multiple-choice item, where the student is asked to select a word from a list (e.g., which word best completes the sentence). Finally, in situations where there are multiple blanks in a paragraph, some tests provide a list of possible words that could fill all blanks in the paragraph (the number of words in this list does not necessarily match the number of blanks in the text). Usually there is only one correct answer per blank (especially in the constrained scenarios), but on occasion, scorers may be provided with a list of acceptable answers per blank.

Location/Identification

These items present students with a visual prompt (e.g., figure, drawing, diagram, graph) in which they are expected to either identify or locate a particular element. For example, an item could provide a map with certain locations marked A through E and a list of five natural resources available in that area; the test-taker is asked to identify which resource is available at a given location. As another example, the item presents test-takers with an empty diagram of an animal's body along with a list of body parts and ask them to locate each body part in the diagram. This item format can become unconstrained (i.e., subjectively-scored) if no options are provided, hence requiring test-takers to recall the terminology required to label the diagram correctly.

Detection and Correction of Errors

These items present the test-taker with a written stimulus (statements, sentences, or paragraphs) which are incorrect according to certain criteria (e.g., veracity or grammar appropriateness) but can be made correct by changing, inserting, or removing one or more words. Sometimes test-takers are only asked to identify the error. More typically, however, they are also required to provide a correction. These items were common in

language arts tests in which test-takers were asked to identify spelling and grammatical errors. Other uses involved judging the veracity of the statements provided and suggesting corrections to correct inaccurate statements.

Problem or Short Answer

This item format provides test-takers with a context and a question to which examinees respond with a short answer. Depending on the content of the test, this item format could require test-takers to answer using one or two sentences or show the steps in a mathematical calculation. This item format is objectively scored when a single correct answer exists or subjectively scored when multiple correct answers exist.

Extended Response

Also referred to as the essay item format, these items require students to write an essay or a response of certain length and complexity. This item type is subjectively scored and often assesses multiple components simultaneously and thus relies on the use of rubrics. This item format has existed since Horace Mann introduced written tests and remains a staple of testing programs.

Technology-enhanced item formats

This section describes the most common technology-enhanced item formats used by large-scale testing programs, namely (a) drag-and-drop, (b) plotting points, (c) selecting text, (d) creating frequency plots, (e) shading areas, (f) creating partitions, (g) hot-spot interactions, (h) matching elements, and (i) in-line drop-down menus. This list was informed by a recent survey of large-scale testing programs that use TEIs conducted by Russell and Moncaleano (2019) and other relevant literature.

Drag-and-Drop

A drag-and-drop item requires the test-taker to select an element presented in the response space, drag it and then drop it in a specific location to produce a response. This interaction is commonly used when the item requires test-takers to classify, order, create, or modify objects. To classify elements, test-takers drag objects provided in the answer space into labeled boxes (e.g., classify a list of animals into vertebrates and invertebrates). Ordering drag-and-drop items ask the test-taker to drag objects into their respective positions in a prescribed order (e.g., ordering the events of a story according to a passage). The drag-and-drop interaction can assist in creating an object, for example, test-takers may create a pictograph by dragging symbols in their respective quantities into an empty set of axes. Finally, this interaction may also be used to modify an object in the answer space, for example, dragging labels into their proper positions in a diagram (e.g., labeling countries in an empty map). The drag-and-drop interaction is not specific to any particular content area.

Plot Points

The plotting points interaction is specific to mathematics and consists of plotting points either in a one-dimensional response space (i.e., a number line) or a two-dimensional response space (i.e., a coordinate grid or blank canvas). The different ways of using this interaction are closely related to the response that is ultimately scored. For some items, the scored response is the position of the points themselves. This occurs, for example, when test-takers are asked to plot a point in a number line or a coordinate grid, where the specific position of the point is the scored response (e.g., identifying the position of a rational number). For other items, the response interaction is used to create a secondary

object, such as a line. In these items, a test-taker may be asked to plot two points that will then be automatically connected to produce a line and the resulting line is the scored response (e.g., plotting a line with a specific slope, the position of the points used may be irrelevant to scoring). Finally, other items require the test-taker to plot multiple points to produce a more complex figure (e.g., graphing a parabola by plotting the vertex and the roots). Some authors often refer to this interaction as "graphing"; however, "plotting points" is a more accurate categorical label as there are items that require plotting points without assessing constructs related to graphing. For example, test-takers may be asked to draw a symmetry line in a figure or create a figure with certain properties (e.g., five-sided convex polygon), neither of which relies on a coordinate grid.

Select Text

The select text interaction is most commonly used in language-related assessments. Items that use this interaction ask test-takers to select a sub-set of text from a larger text. The selection required from a test-taker can vary from single words to sentences or full paragraphs. There is a wide variety of ways in which test-takers may produce a response when engaging with this interaction depending on the delivery system employed. Sometimes test-takers are given a highlighter tool and are allowed to highlight any part of the text. In other cases, the text block is divided into pre-determined clickable sections (e.g., sentences) and the section(s) the test-takers selects constitute their response. Finally, in some cases only a subset of the text is available for selection (e.g., some words throughout the text). The select text interaction is also often referred to as "Hot Text" (Measured Progress, 2016).

Create Frequency Plots

According to Russell & Moncaleano (2019), creating frequency plots is an interaction that requires test-takers to produce histograms, pictograms, or bar graphs. In most cases, this interaction is implemented on a two-dimensional answer space framed by two axes that may be empty or include predefined labels. For histograms and bar graphs, a response may be produced by indicating the height of each bar in the answer space with a mouse click or dragging an existing bar to the desired height. On other occasions, response spaces include "plus" and "minus" buttons below the axis that the test-taker uses to modify the height of the corresponding bar. This latter approach is often used for the production of pictographs, where the buttons are used to add or subtract symbols in the graph. As described earlier, if a delivery system requires test-takers to drag-and-drop symbols to create a pictograph, the item is considered a drag-and-drop interaction item.

Shade Area

This interaction is once again specific to mathematics and involves selecting one or more sections of a figure or area that has been partitioned into a predetermined number of regions. A section of the figure becomes shaded when the user selects it and it is unshaded by clicking on it again. The most common use for this interaction is to produce a visual representation of a fraction. Other uses include selecting the solution region for an inequality in a number line or a system of inequalities in a coordinate grid.

Create Partition

Items that use the create partition interaction require test-takers to divide a region or figure into multiple sections or partitions, often of equal size. To this end, the item provides the test-taker "plus" and "minus" buttons that divide the figure into sections.

The figure is updated automatically as the number of partitions is modified by the test-taker producing equally sized sections (e.g., the first partition of a square would divide it into halves). Items that use this interaction typically require a second action produced through another type of TEI interaction. The most common coupling is with shading area, where test-takers first divide a figure in a desired set of partitions and then shade the sections that will correspond to their response. Another interaction often paired with creating partitions is plotting points, as a test-taker might first be asked to divide a number line (e.g., the segment between 0 and 1) and then plot a point at a specific location (e.g., represent the position of the fraction 2/5).

Hot Spot

This interaction provides a stimulus (an image or a figure) as part of the response space on which test-takers must place a pin or marker to produce a response. The pin can be placed anywhere in the response space by clicking on the desired location. This interaction is used in items that require the test-taker to identify a particular location or element within the stimulus. For example, test-takers might be asked to identify a country in a map or the location of a city within a country's map. Boundaries of regions relevant to the question may or may not be shown for this type of question. Consider the examples mentioned above. In the case of identifying a country, examinees see the boundaries of all countries in the map and may place the pin anywhere within the boundaries of the country they wish to indicate as their response. In contrast, when identifying the location of a city within a country, there is no indication of a region where examinees should place the pin. If region boundaries are not visible to the test-taker, scoring algorithms consider a radius of acceptable answers around the ideal response (usually referred to as tolerance).

Responses within this radius are considered correct. There might also be multiple concentric radii of tolerable responses depending on the purpose of the assessment.

Matching

Items that use the matching interaction present the test-taker with two or more arrays of elements. The test-taker responds by matching elements between these arrays according to criteria included in the prompt. To do so, test-takers first click on an element of one array and then click on an element of a different array. The system then displays a line matching the two clicked elements. Variations to this item format may present arrays of different lengths or include elements in an array that do not have a matching element in another array.

In-Line Choice

Items that employ this interaction present a text (a paragraph or a sentence) where one or more elements (words or phrases) have been replaced by a drop-down menu. Each drop-down menu presents test-takers with multiple options that could complete the text, test-takers are tasked with choosing the most appropriate option. The most appropriate answer may be based on an additional stimulus included in the prompt of the question, such as a reading passage or a diagram. Although some people may not consider this interaction a TEI given the presentation of a constrained set of options, this item format is not replicable in paper-based formats. Moreover, if two or more drop-down menus are presented in the text, the number of possible combinations decreases the chances of guessing the correct answer. TEIs that employ this interaction are also commonly referred to as drop-down menu items.

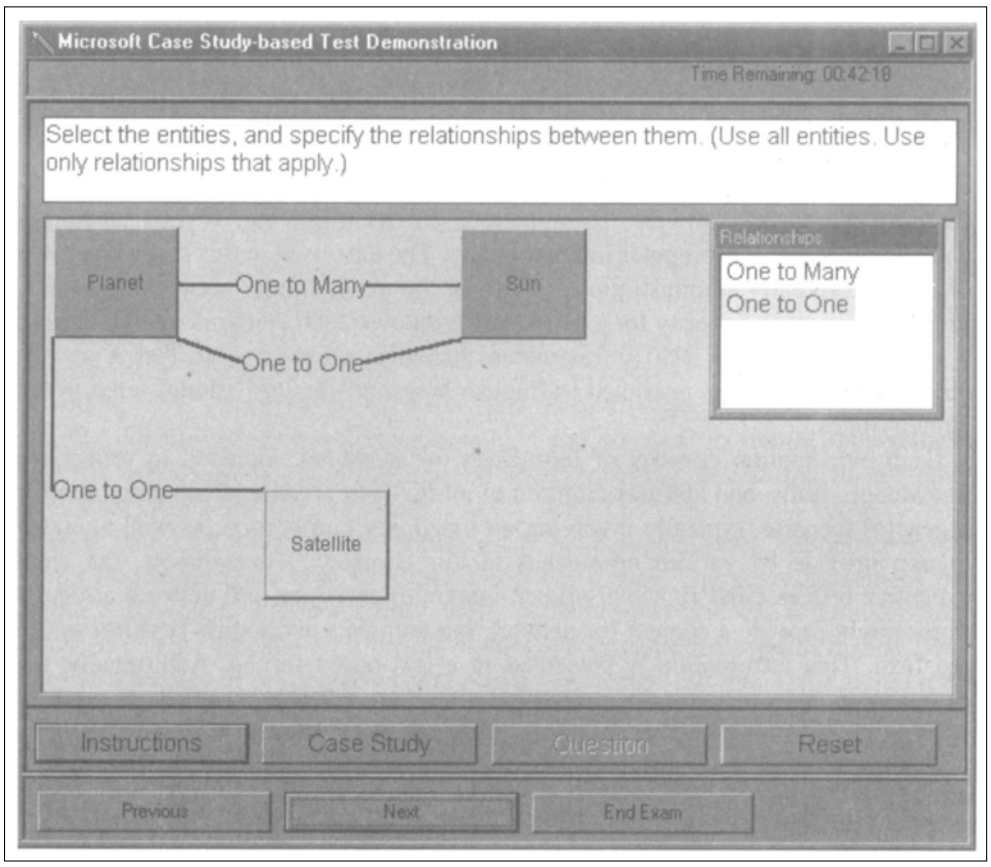
Other Digitally Enhanced Item Types

There are other item formats that have been digitally enhanced and are worth noting given their prevalence in the TEI literature despite not fitting the TEI definition adopted in this work. In some language assessments, the test-takers are allowed to interact with the text in ways that differ from the select text or in-line choice interactions described above. For example, tests may embed blank boxes within a text where test-takers type missing content (fill-in-the-blank items). Other tests provide students with a set of words in a particular place within a text from which the test-taker selects the word that best fits (e.g., conjugations of a verb). For this work, these interactions are considered analogous to traditional items, as they are variations of selected-response or alpha-numeric text-entry interactions.

Innovative item types studied in Jodoin (2003)

Figure A.1

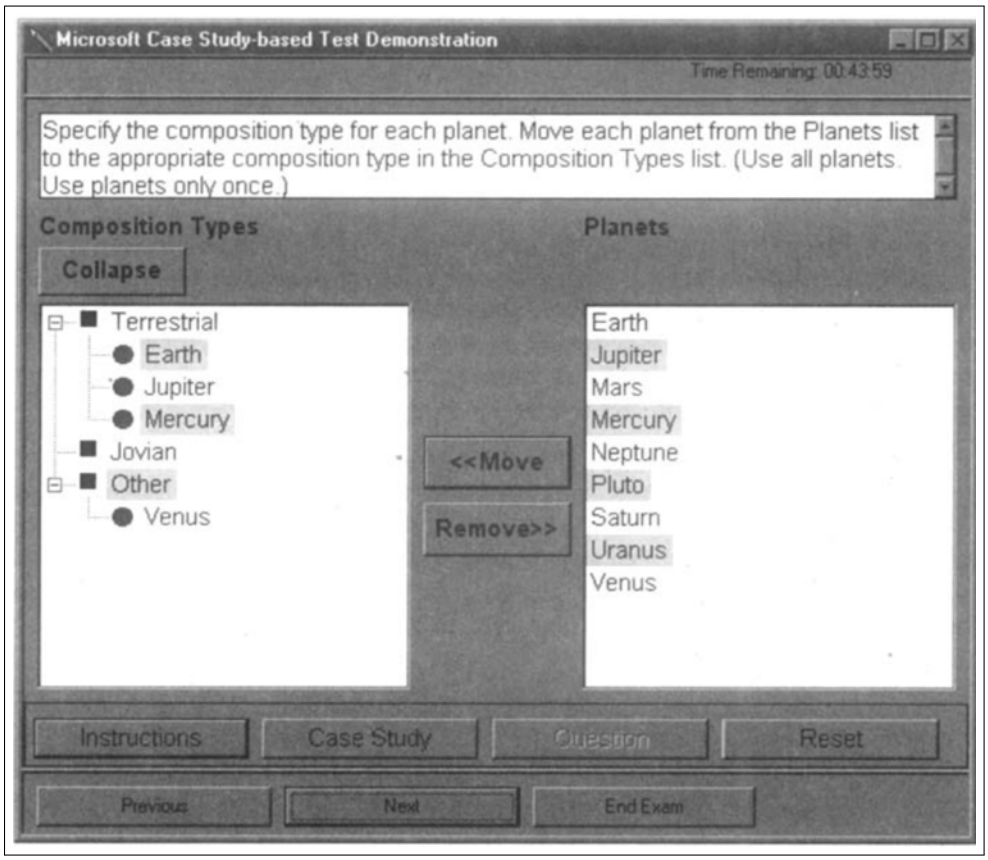
Drop-and-connect (DC) item format



Note. From "Measurement efficiency of innovative item formats in computer-based testing," by Jodoin, M. G., 2003, *Journal of Educational Measurement, 40*(1), p. 4 (https://www.jstor.org/stable/1435051). ©2003 National Council on Measurement in Education.

Figure A.2

Create-a-tree (CT) item format



Note. From "Measurement efficiency of innovative item formats in computer-based testing," by Jodoin, M. G., 2003, *Journal of Educational Measurement, 40*(1), p. 5 (https://www.jstor.org/stable/1435051). ©2003 National Council on Measurement in Education.

Appendix B - Research design and analyses employed in peer-reviewed studies in Rodriguez (2003)

Article	Analysis[a]	Item Equivalence			Research Design		Repeated Measures	
		Stem	Content	None	Within	Between	No	Yes
Ackerman & Smith (1992)	Correlation	X			X			2 weeks
Barnett-Foster & Nagy (1996)	Overall	X				X		
Bennett et al. (1990)	Correlation			X	X			X
Bennett et al. (1991)	Factor Analysis			X	X		X	
Berg & Smith (1994)	Overall	X				X		
Birenbaum & Tatsuoka (1987)	Overall	X				X		
Birenbaum et al. (1992)	Overall	X			X		X	
Bratch & Hopkins (1970)	Correlation			X	X		X	
Breland et al. (1994)	Overall			X	X		X	
Breland & Gaynor (1979)	Correlation			X	X		X	
Bridgeman (1992)	IRT	X				X		
Bridgeman & Lewis (1994)	Overall			X	X		X	
Bridgeman & Rock (1993)	Factor Analysis			X	X		X	
Cirn (1986)	Overall			X	X		X	
Coffman (1966)	Correlation			X	X			
Coulson & Silberman (1960)	ANOVA			X	X		X	
Davis & Fifer (1959)	Correlation	X		X		X		
Eurich (1931)	Correlation		X		X		X	

Article	Analysis[a]	Item Equivalence			Research Design		Repeated Measures	
		Stem	Content	None	Within	Between	No	Yes
Frisbie & Cantor (1995)	Correlation	X				X		
Gay (1980)	ANOVA			X	X		X	
Godshalk et al. (1966)	Correlation			X	X		X	
Halpin et al. (1981)	ANOVA	X				X		
Hancock (1992)	Correlation		X		X		X	
Harke et al. (1972)	Correlation		X		X		X	
Heim & Watts (1967)	Correlation	X			X		X	
Hogan & Mishler (1980)	Correlation			X	X			X
Horn (1966)	Correlation			X	X		X	
Hurd (1932)	Correlation	X			X			1 day
Hurlbut (1954)	Correlation	X			X			1 week
Loyd & Steele (1986)	Correlation			X	X		X	
Lukhele et al. (1994)	IRT			X	X		X	
Magill (1934)	Correlation	X			X		X	
Martinez (1991)	CTT	X				X		
Millman & Setijadi (1966)	Overall	X				X		
Moss et al. (1982)	Correlation			X	X			1 week
Oosterhof & Coats (1984)	CTT	X				X		
Paterson (1926)	Correlation			X		X		
Quelmalz et al. (1982)	Factor Analysis			X	X		X	
Rowley (1974)	Correlation	X			X			5 weeks

Article	Analysis[a]	Item Equivalence			Research Design		Repeated Measures	
		Stem	Content	None	Within	Between	No	Yes
Ruch & Charles(1928)	CTT	X			X			1 day
Ruch & Stoddard (1925)	Correlation	X			X			1 day
Sax & Collet (1968)	ANOVA	X				X		
Thissen et al. (1994)	Factor Analysis			X	X		X	
Traub & Fisher (1977)	Correlation	X			X			2 weeks
van den Bergh (1990)	SEM	X				X		
Vernon (1962)	Correlation		X		X			1-2 weeks
Ward (1982)	Correlation		X		X		X	
Ward et al. (1980)	Factor Analysis		X		X		X	
Wilson & Wang (1995)	IRT			X	X		X	

Note. SEM = Structural Equation Modeling. ANOVA = Analysis of Variance Models. CTT = Evaluation of test and item CTT statistics. IRT = Evaluation of IRT statistics. Overall = Evaluation of overall performance.

Appendix C - Construct fidelity coding guide (Russell & Moncaleano, 2019)

Fidelity	Context	Actions	Examples
High (2)	Authentic (1)	Authentic (1)	**Construct**: Creating graphical representation of a function. **Example**: The interaction presents a coordinate plane on which a student must produce a line that represents the given function. The test taker produces a line by clicking a starting point and dragging the cursor to a second point. The line may be modified by clicking on and moving any point on the line. **Rationale**: Producing a graphical representation of a linear function is a real-world activity. The actions required to produce the line are similar to how lines are produced in Excel, PowerPoint, Word and other applications commonly encountered in the real-world. **Construct**: Apply understanding of fractions. **Example**: The interaction presents the student with tiles of different sizes and requires the student to drag tiles into a response space to create a graphical model of the given fraction. **Rationale**: Creating representations of fractions using tiles is a common classroom activity. The actions required to reposition tiles is similar with how this activity would be done in a digital format during a learning activity. **Construct**: Understanding of positive and negative numbers. **Example**: Student is presented with a number line and asked to shade the section of the number line that satisfies $-3 < x < 5$. **Rationale**: Identifying areas on a number line that satisfy a given condition is a common classroom activity. The actions required to highlight a section of the line are similar to those used by real-world software to highlight content. **Construct**: Identifying text to support a claim. **Example**: A statement is presented and the student is asked to highlight/select text in a passage that supports the statement. The student may select any sentence in the passage by clicking on it or by using a highlighter tool. **Rationale**: Citing text to support a statement, position, or argument is a real-world activity. The actions required to highlight text that supports a statement are similar to how people perform this function using real-world software or when interacting with web-based text.

Fidelity	Context	Actions	Examples
Moderate (1)	Authentic (1)	Inauthentic (0)	**Construct**: Creating graphical representation of a function. **Example**: The interaction presents a coordinate plane and a line with a slope of +1. The student must click and drag the line onto the graph and then rotate the line by clicking on a rotation tool that is located at the top and bottom of the lines to produce a response that reflects the given function. **Rationale**: Producing a graphical representation of a linear function is a real-world activity. The actions required to produce and manipulate the line, however, are not similar to how lines are created or manipulated in software commonly used in the real-world or in a learning environment. **Construct**: Creating functions that represent relationships among variables. **Example**: The student is presented with a table depicting a linear relationship between two variables. The student is presented with "boxes" that represent the elements of a function and a set of numbers and arithmetic symbols. The student is asked to drag the appropriate numbers and symbols into the appropriate boxes to produce a function that represents the relationship between the variables. **Rationale**: When learning to create functions, a common instructional activity involves presenting the structure of the function which the student then completes with numbers are arithmetic symbols; thus this item creates a context similar to one students might experience in the classroom. The act of dragging and dropping numbers and symbols into containers, however, is not similar to the actions students typically take when performing this learning activity. **Construct**: Identifying text to support a claim. **Example**: A statement is presented and the student is asked to select sentences in a passage that supports the statement. A sub-set of sentences is pre-highlighted and only sentences within that sub-set are selectable. **Rationale**: Citing text to support a statement, position, or argument is a real-world activity. In the real-world, however, sentences in a block of text are not pre-highlighted and clicked on to select. For this reason, the actions required to produce a response are not authentic.

Fidelity	Context	Actions	Examples
Low (0)	Inauthentic (0)	Inauthentic (0)	**Construct**: Knowledge of historical events. **Example**: The interaction requires the student to create one or more lines within a defined space that contains a list of historical events on the left side and a list of dates on the right side. **Rationale**: While matching events to dates is a common assessment activity, it is not typically encountered when learning about history and is not an activity in which people engage in the real-world. **Construct**: Ability to list details presented in a block of text. **Example**: The interaction space presents students with a block of text and a separate list of details. Students are required to select those details they believe are related to animals described in the passage and then drag-and-drop those details into a chart labeled "Details to Describe the Animals". **Rationale**: While students are commonly asked to cite details from text, the details are not presented independent of the text. In this example, a list of details is presented from which the student selects details that were contained in a block of text making this an inauthentic task.

Note. From "Examining the Use and Construct Fidelity of Technology-Enhanced Items Employed by K-12 testing Programs," by Russell, M. and Moncaleano, S., 2019, *Educational Assessment, 24*(4), p. 292 (https://doi.org/10.1080/10627197.2019.1670055).

Appendix D - Data Gathering Instruments

This appendix presents the background questionnaire, the data collection instrument, and the closing survey.

Background Questionnaire

Please answer the short survey below before answering the test.

1. How well do you understand English?
   (a) Very well
   (b) Well
   (c) Not very well
2. Have you earned a high school degree?
   (a) Yes
   (b) No
3. Have you earned or are you pursuing higher education a degree related to Mathematics or Statistics?
   (a) Yes
   (b) No, Other (please specify)
   (c) No, I do not have a higher-education degree
4. Have you ever been enrolled in a statistics course at the bachelors or graduate degree level?
   (a) Yes
   (b) No
5. Have you ever taught statistics, mathematics, or science at any academic level before?
   (a) Yes (please specify what subject)
   (b) No

Data Collection Instrument

Five blocks, each with six items, were organized into two forms of the data collection instrument. Both forms presented participants with the common block first followed by blocks TEI-1 and MCI-2 in Form A and blocks MCI-1 and TEI-2 in Form B. Items across blocks TEI-1 and MCI-1 or blocks TEI-2 and MCI-2 were stem-equivalent differing only on the interaction used to produce a response (i.e., drag-and-drop or multiple-choice). Table D.1 shows the order in which items were presented to participants within each block. This section presents blocks and items in the order shown in this table.

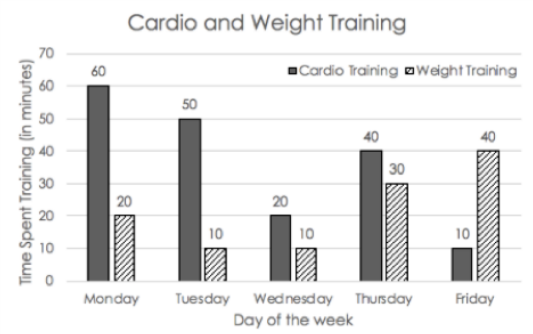
Table D.1

Item order within block

Common	TEI-1	MCI-1	TEI-2	MCI-2
MC1	CL3	MCCL3	CL6	MCCL6
CL1	RO3	MCRO3	RO6	MCRO6
RO1	CL4	MCCL4	CL7	MCCL7
MC2	RO4	MCRO4	RO7	MCRO7
CL2	CL5	MCCL5	CL8	MCCL8
RO2	RO5	MCRO5	RO8	MCRO8

Common Block
CB - MC1

This double bar graph shows the amount of time, in minutes, an athlete spent on cardio training and weight training each day for 5 days.



Which of the following options correctly shows the median amount of time the athlete spent performing each type of training?

Cardio Training	Weight Training
20 minutes	30 minutes

Cardio Training	Weight Training
20 minutes	40 minutes

Cardio Training	Weight Training
25 minutes	35 minutes

Cardio Training	Weight Training
35 minutes	25 minutes

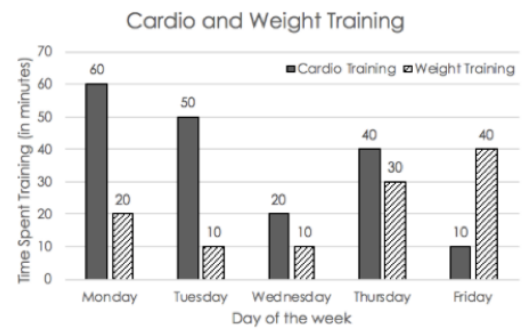
Cardio Training	Weight Training
40 minutes	20 minutes

Cardio Training	Weight Training
40 minutes	30 minutes

Note. Adapted from MCAS Released Items - Mathematics, 2019, Grade 10, Question 14 (Retrieved from http://mcas.pearsonsupport.com/released-items/). ©1998 - 2018 Pearson Education, Inc.

CB - CL1

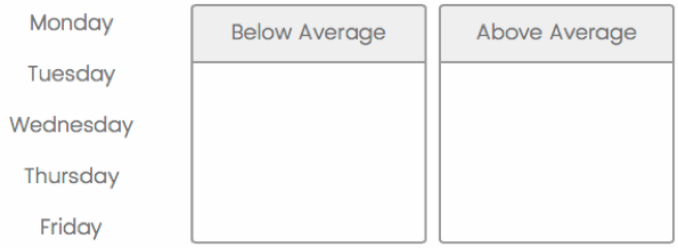
This double bar graph shows the amount of time, in minutes, an athlete spent on cardio training and weight training each day for 5 days.



Calculate the average total amount of time the athlete trained per day.

Classify each day of the week based on whether the athlete spent more or less total time training than the average total training time per day.

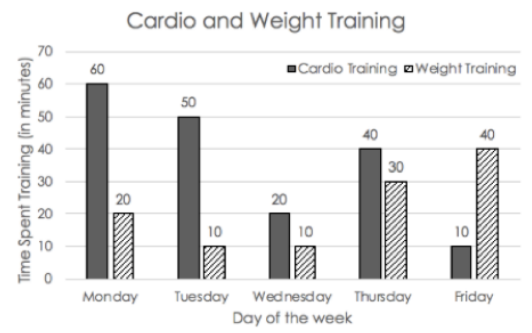
Drag and drop the days of the week that meet these criteria into their corresponding boxes.



Note. Adapted from MCAS Released Items - Mathematics, 2019, Grade 10, Question 14 (Retrieved from http://mcas.pearsonsupport.com/released-items/). ©1998 - 2018 Pearson Education, Inc.

CB - RO1

This double bar graph shows the amounts of time, in minutes, an athlete spent on cardio training and weight training each day for 5 days.



Calculate the total amount of time the athlete trained per day and order the days from **most to least** time spent training.

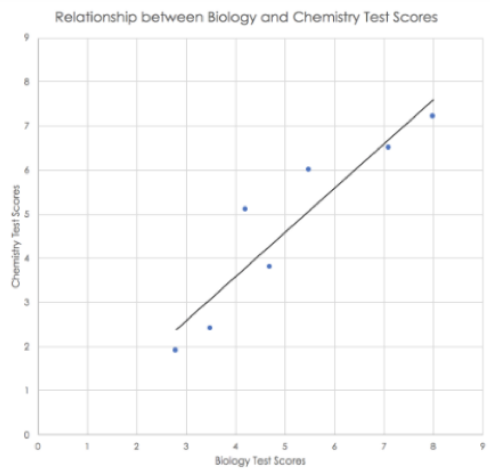
Drag and drop the days of the week in order by time spent training, with the day that the athlete spent the most time training at the top.

- Monday
- Tuesday
- Wednesday
- Thursday
- Friday

Note. Adapted from MCAS Released Items - Mathematics, 2019, Grade 10, Question 14 (Retrieved from http://mcas.pearsonsupport.com/released-items/). ©1998 - 2018 Pearson Education, Inc.

CB - MC2

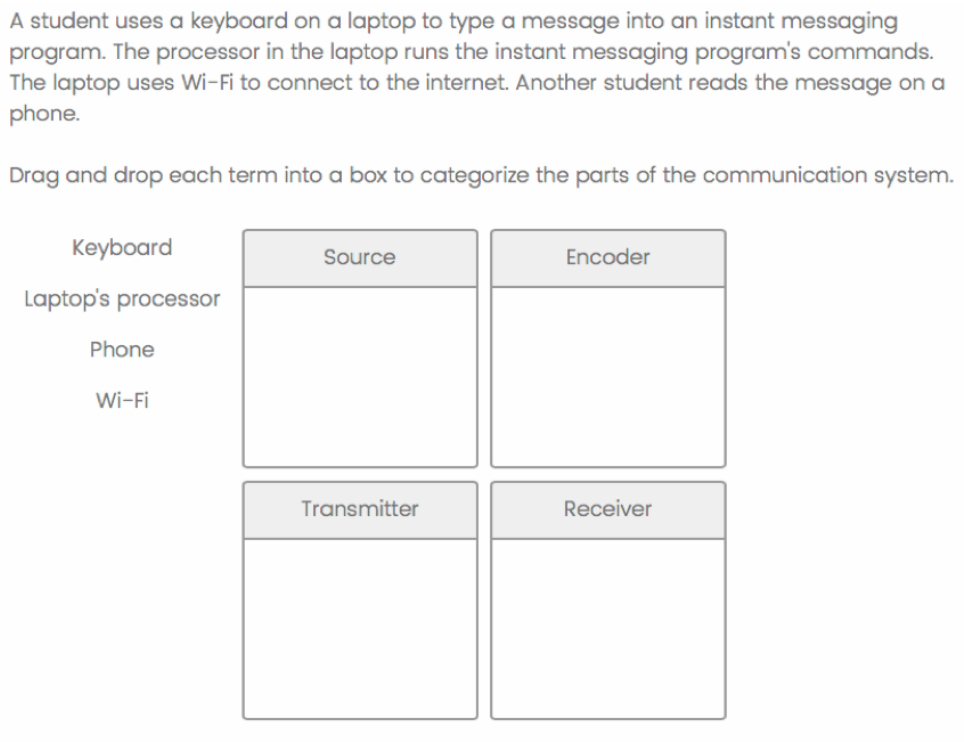
Consider the following scatterplot of the relationship between scores on a biology test and a chemistry test. The dotted line represents a linear regression.



Which of the following options show the predicted score on the chemistry test for a student who scored 5.5 points on the biology test?

- ◯ 4.5
- ◯ 5
- ◯ 5.5
- ◯ 6
- ◯ 6.5
- ◯ 7

CB - CL2

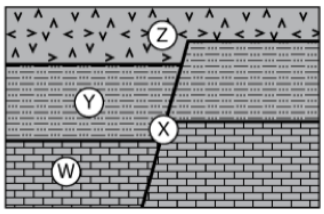


Note. Adapted from MCAS Practice Tests - Mathematics, Grade 8, Question 14 (Retrieved from http://mcas.pearsonsupport.com/student/). ©1998 - 2018 Pearson Education, Inc.

CB - RO2

The diagram shows three rock layers: W, Y, and Z. It also shows fault X.

X, W, Y, and Z were each formed at different times.



Drag and drop the labels to show the order or formation from **oldest to most recent**. Put the label representing the oldest formation on top.

W
X
Y
Z

Note. Adapted from MCAS Practice Tests - Science, Grade 8, Question 17 (Retrieved from http://mcas.pearsonsupport.com/student/). ©1998 - 2018 Pearson Education, Inc.

TEI Block 1

TEI-1 - CL3

Some bodies of water in and around Florida are shown on the map. Some of these bodies of water are freshwater sources, while others are saltwater sources.

Drag and drop each body of water into the correct box to show whether it is a freshwater source or a saltwater source.

Gulf of Mexico

Lake Istokpoga

Kissimmee River

Lake Okeechobee

Atlantic Ocean

Peace River

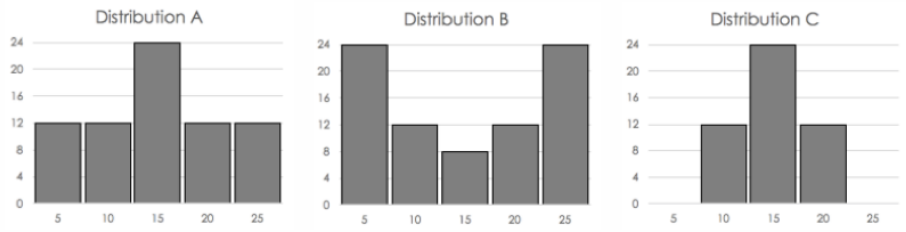
Caloosahatchee River

Freshwater Source	Saltwater Source

Note. Adapted from MCAS Released Items - Science, 2019, Grade 5, Question 18 (Retrieved from http://mcas.pearsonsupport.com/released-items/). ©1998 - 2018 Pearson Education, Inc.

TEI-1 - RO3

The frequency distributions shown below represent three groups of data.



The standard deviation is a measure of the spread of the data and it approximates the average deviation of each score from the mean.

Order the three distributions based on the magnitude of their standard deviations, putting the distribution with the largest standard deviation at the top.

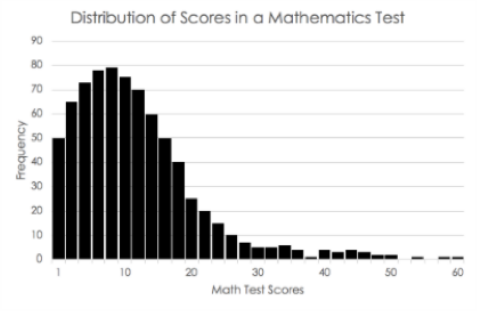
Distribution A

Distribution B

Distribution C

TEI-1 - RO4

This graph shows the frequency distribution of Math scores on the end-of-year state test for a small district.



Arrange the mean, median, and mode of this distribution in order from **largest to smallest**.

Mode

Median

Mean

TEI-1 - CL5

A waiter recorded the amount of money he earned in tips each day for a two-week period. His data are shown in this table.

Week	Monday	Tuesday	Wednesday	Thursday	Friday
1	25	44	48	63	75
2	35	35	48	62	75

Classify each of the following statements according to whether they are true or false.

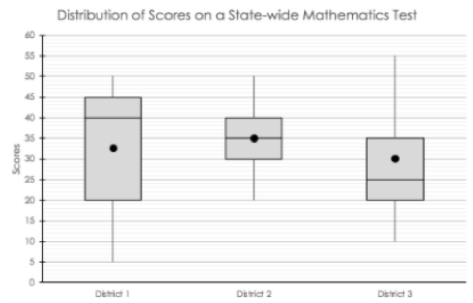
1. The mean and the range for week 1 are equal to the mean and the range for week 2.
2. The mean and the median for week 1 are equal to the mean and the median for week 2.
3. The median and the range for week 1 are equal to the median and range for week 2.

	True	False
Statement 1		
Statement 2		
Statement 3		

Note. Adapted from MCAS Released Items - Mathematics, 2019, Grade 10, Question 10 (Retrieved from http://mcas.pearsonsupport.com/released-items/). ©1998 - 2018 Pearson Education, Inc.

TEI-1 - RO5

The following box plots show the distribution of scores for three districts in a state-wide mathematics test.



Order the districts based on the size of their interquartile ranges from **largest to smallest**, beginning with the district with the largest interquartile range.

District 1

District 2

District 3

MCI Block 1

MCI-1 - MCCL3

Some bodies of water in and around Florida are shown on the map. Some of these bodies of water are freshwater sources, while others are saltwater sources.

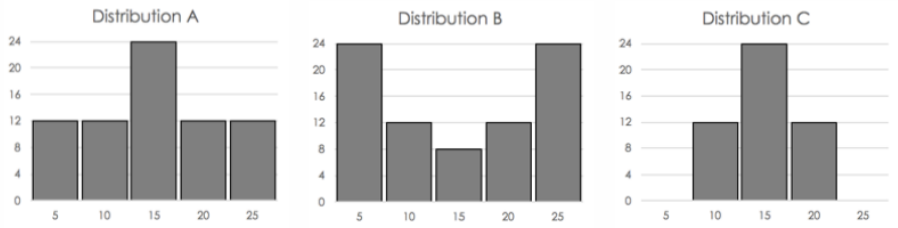
Which of the following correctly lists all the bodies of water that are freshwater sources?

- ◯ Caloosahatchee River, Kissimmee River, Lake Istokpoga, Lake Okeechobee, Peace River
- ◯ Caloosahatchee River, Kissimmee River, Peace River
- ◯ Caloosahatchee River, Kissimmee River, Lake Istokpoga, Peace River
- ◯ Lake Istokpoga, Lake Okeechobee
- ◯ Atlantic Ocean, Gulf of Mexico, Lake Istokpoga, Lake Okeechobee
- ◯ Gulf of Mexico, Caloosahatchee River, Kissimmee River, Peace River

Note. Adapted from MCAS Released Items - Science, Grade 5, Question 18 (Retrieved from http://mcas.pearsonsupport.com/released-items/). ©1998 - 2018 Pearson Education, Inc.

MCI-1 - MCRO3

The frequency distributions shown below represent three groups of data.

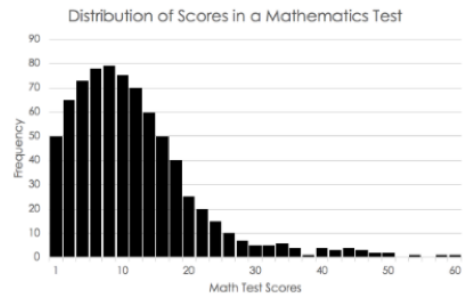


Which of the following statements is **true**?

- The standard deviation of distribution A is the largest while the standard deviation of distribution B is the smallest.
- The standard deviation of distribution B is the largest while the standard deviation of distribution C is the smallest.
- The standard deviation of distribution C is the largest while the standard deviation of distribution A is the smallest.
- The standard deviation of distribution B is the largest while the standard deviation of distribution A is the smallest.
- The standard deviation of distribution A is the largest while the standard deviation of distribution C is the smallest.
- The standard deviation of distribution C is the largest while the standard deviation of distribution B is the smallest.

MCI-1 - MCRO4

This graph shows the frequency distribution of Math scores on the end-of-year state test for a small district.



Which of the following options lists the mean, median, and mode of this distribution in order from **largest to smallest**?

- ◯ Mode, Median, Mean
- ◯ Mode, Mean, Median
- ◯ Median, Mean, Mode
- ◯ Median, Mode, Mean
- ◯ Mean, Median, Mode
- ◯ Mean, Mode, Median

MCI-1 - MCCL5

A waiter recorded the amount of money he earned in tips each day for a two-week period. His data are shown in this table.

Week	Monday	Tuesday	Wednesday	Thursday	Friday
1	25	44	48	63	75
2	35	35	48	62	75

Consider the following statements:

1. The mean and the range for week 1 are equal to the mean and the range for week 2.
2. The mean and the median for week 1 are equal to the mean and the median for week 2.
3. The median and the range for week 1 are equal to the median and the range for week 2.

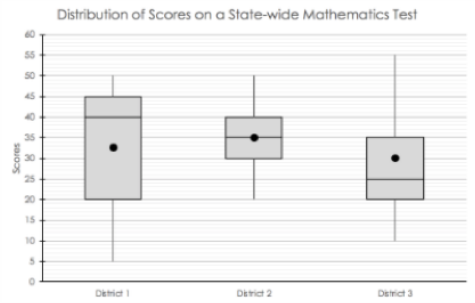
Which of these options lists **all** statements that are true?

- ◯ Only Statement 1
- ◯ Only Statement 2
- ◯ Only Statement 3
- ◯ Statements 1 and 2
- ◯ Statements 1 and 3
- ◯ Statements 2 and 3

Note. Adapted from MCAS Released Items - Mathematics, 2019, Grade 10, Question 10 (Retrieved from http://mcas.pearsonsupport.com/released-items/). ©1998 - 2018 Pearson Education, Inc.

MCI-1 - MCRO5

The following box plots show the distribution of scores for three districts in a state-wide mathematics test.



Which of the following options correctly lists the three districts in order based on the size of the interquartile range from **largest to smallest**?

- ◯ District 1, District 2, District 3
- ◯ District 1, District 3, District 2
- ◯ District 2, District 1, District 3
- ◯ District 2, District 3, District 1
- ◯ District 3, District 1, District 2
- ◯ District 3, District 2, District 1

TEI Block 2

TEI-2 - RO6

A student measured three objects made of different materials. The table below shows each object's mass and volume.

Object	Mass (g)	Volume (cm3)
Block of wood	250	580
Large glass marble	50	21
Rubber ball	500	540

Drag and drop each material in order of density, from **least dense to most dense**, placing the least dense material at the top.

Hint: Density is calculated as the ratio of mass to volume.

Glass

Rubber

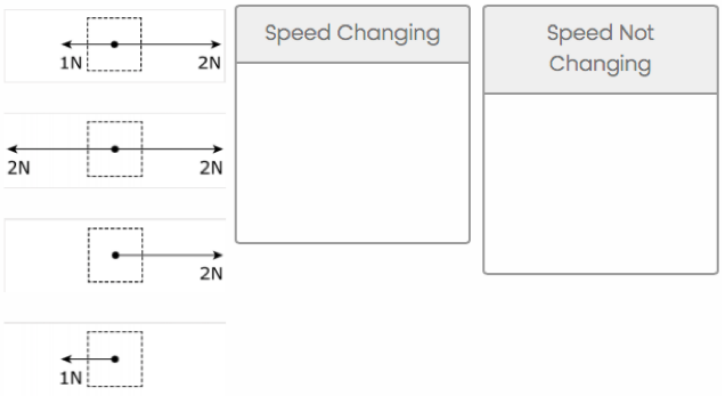
Wood

Note. Adapted from MCAS Released Items - Science, 2019, Grade 8, Question 10 (Retrieved from http://mcas.pearsonsupport.com/released-items/).

TEI-2 - CL7

Four objects have different forces applied to them. All of the horizontal forces acting on each object are shown in the diagrams.

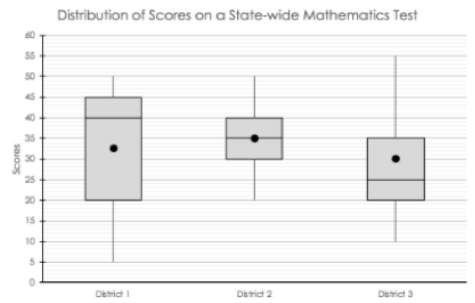
Drag and drop each diagram into a box to show whether the speed of the object is changing.



Note. Adapted from MCAS Released Items - Science, 2019, Grade 8, Question 12 (Retrieved from http://mcas.pearsonsupport.com/released-items/). ©1998 - 2018 Pearson Education, Inc.

TEI-2 - RO7

The following box plots show the distribution of scores for three districts in a state-wide mathematics test.

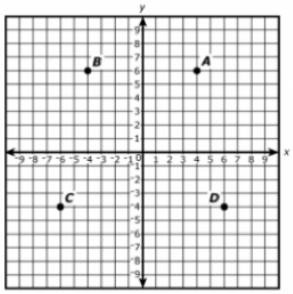


Order the districts based on the mean of the data from **largest to smallest**, placing the district with the largest mean at the top.

- District 1
- District 2
- District 3

TEI-2 - CL8

Four points are shown on this coordinate plane.



Match each ordered pair to its corresponding point on the coordinate plane.

Drag and drop each point into its corresponding box.

(-6 , -4)

(-4 , 6)

(6 , -4)

(4 , 6)

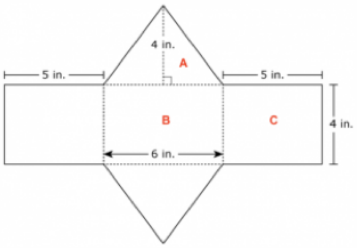
Point A	Point B

Point C	Point D

Note. Adapted from MCAS Released Items - Mathematics, 2019, Grade 6, Question 18 (Retrieved from http://mcas.pearsonsupport.com/released-items/).

TEI-2 - RO8

A gift box is in the shape of a right triangular prism. The diagram of the gift box is shown.



Calculate the perimeter of Triangle A, Rectangle B, and Rectangle C.

Drag and drop each labeled piece of the diagram in order from **the largest to the smallest** perimeter.

Triangle A
Rectangle B
Rectangle C

Note. Adapted from MCAS Released Items - Mathematics, 2019, Grade 6, Question 3 (Retrieved from http://mcas.pearsonsupport.com/released-items/). ©1998 - 2018 Pearson Education, Inc.

MCI Block 2

MCI-2 - MCRO6

A student measured three objects made of different materials. The table below shows each object's mass and volume.

Object	Mass (g)	Volume (cm3)
Block of wood	250	580
Large glass marble	50	21
Rubber ball	500	540

Which of the following correctly lists the materials in order from **least dense to most dense**?

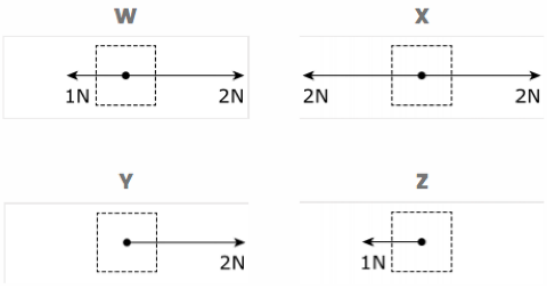
Hint: Density is calculated as the ratio of mass to volume.

- ◯ Glass, Rubber, Wood
- ◯ Glass, Wood, Rubber
- ◯ Rubber, Glass, Wood
- ◯ Rubber, Wood, Glass
- ◯ Wood, Glass, Rubber
- ◯ Wood, Rubber, Glass

Note. Adapted from MCAS Released Items - Science, 2019, Grade 8, Question 10 (Retrieved from http://mcas.pearsonsupport.com/released-items/).

MCI-2 - MCCL7

Four objects labeled W, X, Y, and Z have different forces applied to them. All of the horizontal forces acting on each object are shown in the diagrams.



Which table correctly classifies the four objects according to whether the speeds of the objects are changing?

Speed Changing	Speed Not Changing
Objects Y and Z	Objects W and X

Speed Changing	Speed Not Changing
Objects W and X	Objects Y and Z

Speed Changing	Speed Not Changing
Objects W, Y and Z	Object X

Speed Changing	Speed Not Changing
Object X	Objects W, Y and X

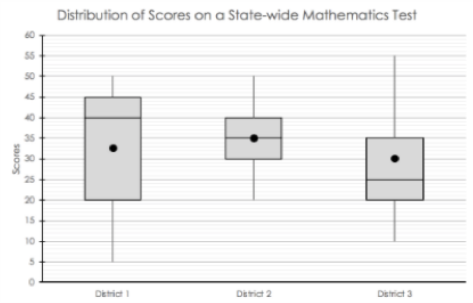
Speed Changing	Speed Not Changing
Objects W and Y	Objects X and Z

Speed Changing	Speed Not Changing
Objects X and Z	Objects W and Y

Note. Adapted from MCAS Released Items - Science, 2019, Grade 8, Question 12 (Retrieved from http://mcas.pearsonsupport.com/released-items/). ©1998 - 2018 Pearson Education, Inc.

MCI-2 - MCRO7

The following box plots show the distribution of scores for three districts in a state-wide mathematics test.

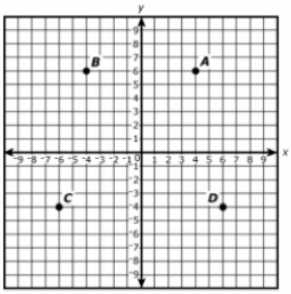


Which of the following options correctly lists the three districts in order based on their means from **largest to smallest**?

- ◯ District 1, District 2, District 3
- ◯ District 1, District 3, District 2
- ◯ District 2, District 1, District 3
- ◯ District 2, District 3, District 1
- ◯ District 3, District 1, District 2
- ◯ District 3, District 2, District 1

MCI-2 - MCCL8

Four points are shown on this coordinate plane.



Which of the following correctly shows the coordinates of each of the points?

Point A	Point B	Point C	Point D
(6,4)	(6,-4)	(-4,-6)	(-4,6)

Point A	Point B	Point C	Point D
(6,4)	(-4,6)	(-4,-6)	(6,-4)

Point A	Point B	Point C	Point D
(6,4)	(6,-4)	(-6,-4)	(6,-4)

Point A	Point B	Point C	Point D
(4,6)	(-4,6)	(-6,-4)	(6,-4)

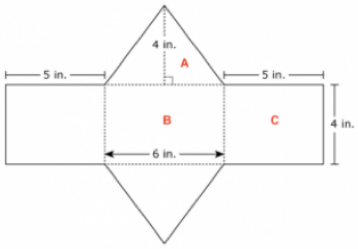
Point A	Point B	Point C	Point D
(4,6)	(-4,6)	(6,-4)	(-6,-4)

Point A	Point B	Point C	Point D
(4,6)	(6,-4)	(-4,-6)	(-4,6)

Note. Adapted from MCAS Released Items - Mathematics, 2019, Grade 6, Question 18 (Retrieved from http://mcas.pearsonsupport.com/released-items/).

MCI-2 - MCRO8

A gift box is in the shape of a right triangular prism. The net of the gift box is shown.



Calculate the perimeter of Triangle A, Rectangle B, and Rectangle C.

Which of the following options correctly lists the labeled pieces of the net from **largest to smallest** perimeter?

- ◯ Triangle A, Rectangle B, Rectangle C
- ◯ Triangle A, Rectangle C, Rectangle B
- ◯ Rectangle B, Triangle A, Rectangle C
- ◯ Rectangle B, Rectangle C, Triangle A
- ◯ Rectangle C, Rectangle B, Triangle A
- ◯ Rectangle C, Triangle A, Rectangle B

Note. Adapted from MCAS Released Items - Mathematics, 2019, Grade 6, Question 3 (Retrieved from http://mcas.pearsonsupport.com/released-items/). ©1998 - 2018 Pearson Education, Inc.

Closing Survey

1. Please describe your experience with statistics. For example, have you taken any statistics courses? if so, at what level? (High school, college, graduate school). Do you work with statistics as part of your job?

2. Did you take any breaks or were interrupted while taking the test?

   (a) Yes

   (b) No

3. How much time do you estimate you spent working on the test?

   (a) Less than 15 minutes

   (b) Between 15 minutes and 30 minutes

   (c) Between 30 minutes and 45 minutes

   (d) Between 45 minutes and 1 hour

   (e) More than 1 hour

Appendix E - Path mapping of the CMV Protocol