

Evaluating the Validity of the eTIMSS 2019 Mathematics Problem Solving and Inquiry Tasks:

Author: Kerry Cotter

Persistent link: <http://hdl.handle.net/2345/bc-ir:108388>

This work is posted on [eScholarship@BC](#),
Boston College University Libraries.

Boston College Electronic Thesis or Dissertation, 2019

Copyright is held by the author. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License ([http://
creativecommons.org/licenses/by-nc-sa/4.0](http://creativecommons.org/licenses/by-nc-sa/4.0)).

Boston College
Lynch School of Education

Department of
Measurement, Evaluation, Statistics, and Assessment

**EVALUATING THE VALIDITY OF THE
eTIMSS 2019 MATHEMATICS
PROBLEM SOLVING AND INQUIRY TASKS**

Dissertation

by

KERRY COTTER

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

May 2019

Evaluating the Validity of the eTIMSS 2019 Mathematics Problem Solving and Inquiry Tasks

By

Kerry Cotter

Dissertation Director: Ina V.S. Mullis

Abstract

The eTIMSS mathematics PSIs were a new and pioneering effort to capitalize on the computer- and tablet-based mode of assessment delivery introduced in the eTIMSS 2019 assessments at the fourth and eighth grades. The PSIs were scenario-based mathematics problem solving tasks intended to enhance measurement of mathematics problem solving and reasoning skills and increase student engagement and motivation in the assessment. These unique tasks were designed to measure the same mathematics content as the rest of the mathematics items in the eTIMSS 2019 assessments, but because of their novelty, there was a question about whether the PSIs achieved this goal and could be reported together with the regular TIMSS mathematics items.

Following a full-scale field test in 30 countries completed in May 2018, this dissertation conducted an in-depth investigation of the validity of the eTIMSS 2019 mathematics PSIs with the goals of informing analysis and reporting plans for TIMSS 2019 and providing insights for future assessments aspiring to capitalize on digital technology. This investigation involved three key tasks: 1) examining and documenting the methods and procedures used to develop the PSIs and promote validity by design, 2) investigating the characteristics of the PSIs in terms of the content coverage and fidelity of student responses, and 3) using the eTIMSS field test data to evaluate the internal structure of the PSIs.

The results indicate that the PSIs are well-aligned with the *TIMSS 2019 Mathematics Framework* and elicited the intended interactions from students. The regular and PSI items were found to measure the same unidimensional construct, and therefore can be validly reported together on the TIMSS 2019 achievement scale. The lessons TIMSS learned in developing the PSIs for eTIMSS 2019 and suggestions for the future also are discussed.

Acknowledgements

First and foremost, thank you to my dissertation advisor and mentor, Dr. Ina Mullis, and reader Dr. Michael Martin, for encouraging me to pursue a doctoral degree, teaching me so much about international large-scale assessment, and helping to make this dissertation possible. I am extremely grateful for your continued support and the opportunity to work under your guidance at the TIMSS & PIRLS International Study Center.

Thank you to Dr. Zhushan “Mandy” Li and Dr. Mary Lindquist for serving as readers on my dissertation committee. Your unique expertise in psychometrics and mathematics education, respectively, were critical in making this dissertation a success. I am very appreciative of your thoughtful suggestions.

Dr. Bethany Fishbein, thank you for the constant reassurance and counsel on all things dissertation related, at any time of the day. I am so fortunate to have you as an officemate and friend.

Thank you to all of the staff at the TIMSS & PIRLS International Study Center. In particular, thank you to Steven Simpson and Yeny Pardini for your work on graphics and programming for the PSIs and to Pierre Foy and Dr. Joe Galia for your help with the field test data. Thank you also to Erin Wry, for keeping me company on late nights and weekends at the office, and to Victoria Centurino, for always offering help and cheeriness.

Thank you to my family and friends for your support, encouragement, and understanding through this challenging endeavor. Primarily, thank you to my parents (Kathryn and John) for always believing in me, offering prudent advice, and inspiring me to have confidence in my academic abilities. Thank you to Lauren Wilkes, for keeping my spirits up and balancing out the stress level in our apartment, and to Tom Vieth, for listening to me constantly talk about my dissertation and celebrating every milestone along the way.

Finally, thank you to all of the individuals who facilitated TIMSS’ ambitious transition to digital assessment—the staff at IEA Hamburg and IEA Amsterdam, the TIMSS 2019 National Research Coordinators, and the schools, students, and teachers that participated in the eTIMSS 2019 Field Test.

Table of Contents

Chapter 1: Introduction	1
The TIMSS Mathematics Assessments.....	2
TIMSS’ Transition to eTIMSS.....	4
The eTIMSS Mathematics PSIs	6
Assessment Validity	8
Dissertation Goals	10
Chapter 2: Literature Review.....	14
Mathematics Education and Large-Scale Assessment	14
Benefits of eAssessment.....	18
Construct Validity.....	19
Student Engagement.....	21
Operational Efficiency.....	22
Challenges in Developing the eTIMSS Mathematics PSIs	23
Adhering to Best Practices in Item Writing.....	23
Assessing Problem Solving in Mathematics	24
Developing Problem Contexts.....	26
Assessing Mathematics on a Computer.....	27
Examples of Large-Scale Problem Solving Assessments	29
NAEP – Problem Solving in Technology-Rich Environments	29
PISA – Complex and Collaborative Problem Solving	30
Benefits and Challenges of Complex Problem Solving Tasks	31
“Beyond Constructed Response Items”.....	33
Summary of Benefits and Challenges of eAssessment	34
Promoting and Demonstrating Assessment Validity	35
Validity Evidence Based on Test Content.....	37
Validity Evidence Based on Response Process	39
Validity Evidence Based on Internal Structure	44
Chapter 3: Methods	53
PSI Development Methods to Promote Validity.....	54

Overview	54
Developing the TIMSS 2019 Mathematics Framework.....	55
Initial Task Development	57
Expert Review	58
eAssessment Systems and Programming	59
Cognitive Laboratories, Pilot Testing, and Observations.....	64
Developing Scoring Guides for Constructed Response Items.....	75
eTIMSS 2019 Field Test	77
Overview	77
Achievement Instrument Design	77
eTIMSS Student Questionnaire	79
Sample	80
Field Test Data Collection.....	82
Scoring Constructed Response Items	83
Feedback from NRCs	85
Item Review.....	86
Analysis Methods.....	87
Overview	87
Validity Evidence Based on Test Content.....	88
Validity Evidence Based on Response Process.....	91
Validity Evidence Based on Internal Structure	94
Chapter 4: Results.....	118
Test Content Validity	119
eTIMSS Mathematics Framework Coverage with the PSIs.....	120
Validity of Response Process.....	126
Functionality and Usability	126
Students Found eTIMSS Engaging	143
PSI Items were Scored Reliably	148
Validity of Internal Structure	149
Speededness and Position Effects.....	150
Measurement Properties of Items	154

Performance Consistency across Regular and PSI Items	156
Underlying Factor Structure	158
Summary of Results	167
Chapter 5: Discussion	170
Summary of Key Findings	172
Lessons Learned	174
Suggestions for the Future	178
Implications for the Future of the PSIs	181
References	183
Appendix	198
Appendix A: eTIMSS Mathematics PSI Development Milestones	198
Appendix B: Selected Survey Activities Questionnaire Items.....	202
Appendix C: eTIMSS Student Questionnaire Results	203
Appendix D: Average Time per Screen for the Mathematics PSIs in the eTIMSS Field Test	204

List of Exhibits

Exhibit 3.1: eTIMSS Assessment System.....	60
Exhibit 3.2: eTIMSS Player Interface.....	62
Exhibit 3.3: eTIMSS 2019 prePilot Blocks	69
Exhibit 3.4: eTIMSS 2019 prePilot Block Combinations.....	70
Exhibit 3.5: eTIMSS 2019 Field Test Block Combinations/Booklets.....	78
Exhibit 3.6: eTIMSS Student Questionnaire Items Measuring Student Enjoyment and Difficulties Taking the Test on a Computer or Tablet	80
Exhibit 3.7: Countries in the eTIMSS 2019 Field Test.....	81
Exhibit 3.8: Summary of eTIMSS 2019 Field Test Items and Participants.....	82
Exhibit 3.9: Number of Items and Responses from the eTIMSS 2019 Field Test Used in Analysis.....	97
Exhibit 3.10: eTIMSS 2019 Field Test Data Matrix.....	101
Exhibit 3.11: Analysis Models Used to Investigate the Underlying Structure of the eTIMSS 2019 Mathematics Field Test Data – Class-level Analysis	111
Exhibit 3.12: Analysis Models Used to Investigate the Underlying Structure of the eTIMSS 2019 Field Test Data – Student-level Analysis Including Science Items....	113
Exhibit 4.1: Fourth Grade eTIMSS 2019 Mathematics PSI Problem Scenarios and Framework Content Domain Topics within Number, Measurement and Geometry, and Data	119
Exhibit 4.2: Eighth Grade eTIMSS 2019 Mathematics PSI Problem Scenarios and Framework Content Domain Topics within Number, Algebra, Geometry, and Data and Probability	120
Exhibit 4.3: eTIMSS 2019 Mathematics Assessments by Content Domain.....	122
Exhibit 4.4: eTIMSS 2019 Mathematics Assessments by Cognitive Domain	124
Exhibit 4.5: International Average Item Statistics for Mathematics Ruler Tool Items in the eTIMSS 2019 Field Test	130
Exhibit 4.6: Number of Regular and PSI Mathematics Items in the eTIMSS 2019 Field Test by Item Type	132
Exhibit 4.7: International Average Item Statistics from the Regular Mathematics Items in the eTIMSS/paperTIMSS 2019 Field Test by eTIMSS Item Type and Mode of Administration.....	134

Exhibit 4.8: Functionality of the eTIMSS Number Pad for Numeric Constructed Response Items	136
Exhibit 4.9: Average Time per Screen in the eTIMSS 2019 Field Test by Cognitive Domain and Score Points – Grade 4	145
Exhibit 4.10: Average Time per Screen in the eTIMSS 2019 Field Test by Cognitive Domain and Score Points – Grade 8	146
Exhibit 4.11: Machine- and Human-Scored Constructed Response Items in Mathematics PSIs in the eTIMSS 2019 Field Test	148
Exhibit 4.12: International Average Time per Block in the eTIMSS 2019 Field Test..	151
Exhibit 4.13: International Average Item Statistics from the eTIMSS 2019 Field Test for Mathematics Items Selected for eTIMSS 2019 Data Collection.....	155
Exhibit 4.14: Average Percent Correct on Regular versus PSI Items by Country for the eTIMSS Mathematics Items Selected for eTIMSS 2019 Data Collection.....	157
Exhibit 4.15: Number of Parameters, Dimensions, Deviance, AIC, and BIC for Class-Level Analysis Models.....	158
Exhibit 4.16: Median Standardized Factor Loadings for Class-Level Analysis Models	160
Exhibit 4.17: Number of Parameters, Dimensions, Deviance, AIC, and BIC for Student-Level Analysis Models	162
Exhibit 4.18: Median Standardized Factor Loadings for Student-Level Analysis Models.....	162
Exhibit 4.19: Plots of Standardized Factor Loadings on General versus Specific Factors from Bi-factor Models – Grade 4.....	164
Exhibit 4.20: Plots of Standardized Factor Loadings on General versus Specific Factors from Bi-factor Models – Grade 8.....	166
Exhibit A.1: eTIMSS Mathematics PSI Development Milestones, January 2015–September 2018.....	198
Exhibit C.1: International Summary Statistics for Students Like Taking the Test on a Computer or Tablet	203
Exhibit C.2: International Summary Statistics for Students Experiencing Difficulties Taking eTIMSS.....	203
Exhibit D.1: Average Time per Screen in the eTIMSS Field Test – Grade 4 PSI Blocks.....	204
Exhibit D.2: Average Time per Screen in the eTIMSS Field Test – Grade 8 PSI Blocks.....	205

Chapter 1: Introduction

TIMSS (Trends in International Mathematics and Science Study) is a large-scale international assessment of student achievement in mathematics and science at the fourth and eighth grades that has been conducted every four years since 1995. The TIMSS assessments are curriculum-based and designed to measure the content and cognitive dimensions delineated in frameworks developed collaboratively with participating countries and updated for each assessment cycle. The most recent mathematics and science frameworks are found in the *TIMSS 2019 Assessment Frameworks* (Mullis & Martin, 2017).

TIMSS data are collected in more than 60 countries, with many of them having trend data back to the first assessment. For more than 20 years, TIMSS has provided valid and reliable measurement of student achievement, supporting participating countries in measuring the effectiveness of their education systems in a global context, monitoring the impact of educational initiatives, and stimulating curriculum reform (Mullis, Martin, Goh & Cotter, 2016). Educators and measurement specialists around the world also look to TIMSS as an exemplar of high quality assessment and commonly use the TIMSS assessment frameworks and achievement items to inform the development of national and regional examinations as well as train teachers in measuring student achievement (Mullis et al., 2016).

The TIMSS assessments are directed by IEA's TIMSS & PIRLS International Study Center at Boston College. IEA (International Association for the Evaluation of Educational Achievement) is an independent international cooperative of national

research institutions and government agencies that pioneered studies of cross-national achievement in the 1950s. IEA is headquartered in Amsterdam and Hamburg.

TIMSS assesses student achievement in both mathematics and science at the fourth and eighth grades, but this dissertation primarily focuses on the mathematics assessments.

The TIMSS Mathematics Assessments

The TIMSS mathematics assessments are designed to provide internationally comparable student achievement results on the mathematics content and skills that are valued by the international mathematics education community and included in the curricula of participating countries. Mathematics educators around the world strongly believe that in addition to teaching students mathematical facts and principals, it is vital that students are prepared to use the mathematics learned in the classroom to solve problems in the real world (e.g., Boaler, 1993; Darling-Hammond et al., 2013; Husen, 1967; Kilpatrick, 1992; Polya, 1957; Schoenfeld, 1985). Mathematics is a part of a variety of daily life activities, such as managing money, cooking, and building things. Many career fields, including medicine, computer science, engineering, and business, require a deep understanding of mathematics for success (Lindquist, Philpot, Mullis & Cotter, 2017).

Therefore, acquiring the mathematical practices needed to use mathematics beyond the classroom is considered as much a part of becoming mathematically literate as any other defined content standard (Scherrer, 2015, p.199). In the real world, mathematics problems are complex, unfamiliar, and require multiple steps—the

mathematics cannot be “decoupled” from the situation (Scherrer, 2015). For this reason, it is essential that students learn and are assessed on both the mathematics content covered in the curriculum and the processes and procedures needed to apply mathematical knowledge beyond the classroom to promote mathematical literacy. Given the symbiotic relationship between what is tested and what is taught, it is particularly important that large-scale assessments address these practices (Barnes, Clarke & Stephens, 2010; Swan & Burkhardt, 2012).

To support the critical purpose of mathematics education to develop students’ problem solving skills and to provide valid and reliable measurement of mathematics ability, mathematics assessments must include items that span the full spectrum of cognitive demands (Lindquist et al., 2017; Pellegrino, Chudowsky & Glaser, 2001; Suurtamm et al., 2016; Swan & Burkhardt, 2012). To achieve this goal, the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017) is organized around two dimensions: the content domains, specifying the subject matter to be assessed and the cognitive domains, specifying the thinking processes to be assessed as students engage with the content. At both grades, there are three mathematics cognitive domains—knowing, applying, and reasoning. *Knowing* is the most basic, covering the mathematics facts, concepts, and procedures students need to know. *Applying* goes a step further, focusing on students’ ability to apply knowledge and conceptual understanding to solve problems or answer questions, and *reasoning* goes beyond solving routine problems to encompass unfamiliar situations, complex contexts, and multi-step problems.

To reflect the priorities of the international mathematics education community and country’s intended mathematics curricula, TIMSS continues to devote 60 percent of the fourth grade mathematics assessment and 65 percent of the eighth grade mathematics assessment to measuring applying and reasoning skills.

TIMSS’ Transition to eTIMSS

For the 2019 assessment cycle, TIMSS transitioned to eTIMSS—an electronic version of the TIMSS assessments designed for computer- and tablet-based administration through an eAssessment system developed by IEA Hamburg in collaboration with the TIMSS & PIRLS International Study Center. More than half the approximately 60 countries participating in TIMSS 2019 elected to administer the “e” version of the assessment. Data collection began in the Southern Hemisphere in September 2018 and continued in the Northern Hemisphere through June 2019.

Transitioning to digital assessment is important to “keep up with the times” and is expected to increase construct representation and data utility (Bennett, 2015; Braun, 2013; O’Leary, Scully, Karakolidis & Pitsia, 2018). eAssessment offers a wider variety of item types that may be well-suited for assessing complex areas of the framework that were historically challenging to measure using paper-based assessment, such as mathematics problem solving and reasoning (NCTM Research Committee, 2013; Scalise, 2012). Computer- and tablet-based delivery can also be more interactive and engaging for students who are increasingly accustomed to learning on a computer (Bennett, 2015; TIMSS & PIRLS International Study Center, 2017). Further, eAssessment provides the benefits of increased operational efficiency in translation, assessment delivery, data entry,

and scoring, which are particularly important for a large-scale international project (Poggio, Glasnapp, Yang & Poggio, 2005).

The TIMSS design requires keeping more than half the assessment items (four-sevenths) secure from cycle to cycle to measure trends, and the rest are newly developed for each assessment cycle. The TIMSS 2019 trend items from the previous assessment in 2015 were converted to digital format and the new items for the 2019 cycle were designed to be administered via the eAssessment system. The eTIMSS countries also administered the trend items in paper format as a bridge from 2015. The current plan is to use this bridge to enable reporting the eTIMSS and paperTIMSS results in 2019 on the same achievement scale.

To capitalize on the superior design features eAssessment offers and improve measurement of higher-order mathematics skills, TIMSS 2019 went beyond developing traditional TIMSS items in digital format. Development work also included an ambitious initiative to create a series of extended Problem Solving and Inquiry (PSI) tasks. The goal of the PSIs was to measure student achievement in mathematics and science in a more authentic way than is possible with traditional paper-and-pencil achievement items. Historically, TIMSS 2003 was funded by the National Science Foundation (NSF) to develop and assess such longer problem solving and inquiry tasks. However, in the paper format students became either “lost” or “bored” by these extended tasks and the participating countries asked that they be discontinued. TIMSS thought such extended tasks could be successful in the computerized format where students would be more engaged, motivated, and supported as necessary to complete the tasks.

The eTIMSS Mathematics PSIs

The mathematics PSIs were designed for eTIMSS 2019 with the goal of improving coverage of the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017) by enhancing measurement of students' applying and reasoning skills across the mathematics content domains. Each mathematics PSI consists of a sequence of 4 to 12 items that are set in a cohesive context and address a range of topics in the *TIMSS 2019 Mathematics Framework* (e.g., building a shed to store equipment or adding information to a website by solving a series of mathematics problems). The items within these situational tasks take advantage of technology by including animations, colorful graphics, and interactive response spaces. The items guide students through problem scenarios to provide scaffolding for complex mathematics problems that would be difficult to ask without the appropriate support.

Anchored in the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017), the PSIs are designed to measure subject-specific problem solving ability in mathematics, rather than domain-general skills. This differentiates the PSIs from other digital large-scale assessments of problem solving that assess problem solving as a 21st century skill (e.g., Programme for International Student Assessment's (PISA) Complex and Collaborative Problem Solving assessments and National Assessment of Educational Progress' (NAEP) Technology and Engineering Literacy assessment).

From the onset, the PSIs were seen as a chance to realize the potential of eAssessment and were afforded a broader array of innovative digital features beyond

those available for regular eTIMSS items. In the early stages of developing the PSIs, TIMSS established four criteria for a mathematics PSI:

- 1) Assess mathematics problem solving (not primarily reading or perseverance);
- 2) Take advantage of the “e” environment;
- 3) Be engaging and motivating for students;
- 4) Be administered and scored via the TIMSS eAssessment systems.

Developing an eAssessment system and technology-enhanced achievement items is a substantial undertaking in and of itself, and given their complexity, the eTIMSS PSIs were especially challenging and resource intensive to develop. During the nearly four-year development period, the PSIs underwent numerous iterations and changed considerably from their inception based on feedback from mathematics content experts, measurement specialists, developers with expertise in interface design, results of cognitive laboratories, pilot testing, and technical constraints. Through this extended development process, TIMSS learned a number of valuable lessons about the complexities of leveraging technology to assess mathematics problem solving skills.

The mathematics PSIs are a unique and somewhat experimental addition to the eTIMSS 2019 mathematics assessments at the fourth and eighth grades with potentially different measurement properties from the more traditional eTIMSS items. Because the PSIs are a separate effort only applicable to eTIMSS and not to paperTIMSS, there is a question at this point about whether the PSIs extend the TIMSS 2019 mathematics achievement scale or are a different construct. Should the PSIs be included in the TIMSS mathematics scale for reporting trends in TIMSS 2019? Should they be included as an

integral part of the TIMSS 2023 assessments and beyond? This dissertation provides data to help answer these questions as well as insights into the complexities inherent in developing the digital assessments of the future.

Assessment Validity

When developing an assessment it is critical to demonstrate that it measures what it purports to measure, or establish its validity for the intended use. Current validity theory regards test validity as a unified concept, which can be established by gathering and synthesizing a variety of evidence to produce a coherent argument in support of the proposed interpretations and uses of test scores (Kane, 2006; Messick, 1989, 1990). Validity arguments are built within a network of assumptions, through which the observed scores on an instrument can be connected to the conclusions and decisions made about the observed scores by using detailed statements for how observations can be interpreted and specifications for how the interpretations can be evaluated (Kane, 2001). The plausibility of the proposed interpretations are evaluated through the validity argument, which critically examines the inferences and assumptions on which the interpretation is based (Kane, 2001).

A sound validity argument requires evidence based on test content, response process, internal structure, relationships with other variables, and consequences of the test (American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) [AERA, APA & NCME], 2014). According to the *Standards for Educational and Psychological Testing* (2014), “there is always more information that can be gathered to

more fully understand a test and the inferences that can be drawn from it,” and the amount of information needed may vary based on the context of the test and the importance of the desired inferences (p. 22). Collecting validity evidence is seen as an ongoing process, which should begin with defining the construct the test will measure and continue through score reporting and the decisions made based on the results.

Countries around the world use TIMSS results to inform decisions in education policy, making it essential to ensure that the TIMSS assessments are of the highest quality so that decisions based on TIMSS scores are valid. Large-scale assessments such as TIMSS also present a view of the knowledge and skills that are valued by the educational community, so it is crucial that the content, processes, and types of tasks used in TIMSS are representative of the mathematics that is most important to educators (Davis, 1992; Suurtamm et al., 2016; Swan & Burkhardt, 2012). The achievement items used in large-scale assessments influence teachers’ instruction and assessment practices, and consequently students’ achievement and perceptions of learning mathematics (Suurtamm et al., 2016; Swan & Burkhardt, 2012).

As demonstrated by TIMSS’ methods and procedures documentation from previous assessment cycles (e.g., Martin & Mullis, 2012; Martin, Mullis & Hooper, 2016), TIMSS has a longstanding history of adhering to best practices in instrument development and transparent reporting of the actions taken to promote validity throughout the assessment process. Continuing this practice, *TIMSS Methods and Procedures in 2019* (forthcoming) will describe TIMSS’ end-to-end process for the entire assessment. However, given the novelty and complexity of the eTIMSS PSIs, their

development followed a slightly different trajectory than the regular eTIMSS items, necessitating increased attention to their validity in development and in considering analysis and reporting plans. In the midst of the transition to digital assessment, it is especially important to critically evaluate each new innovation in the field to support progress toward realizing the full potential of technology-enhanced assessment (O’Leary et al., 2018).

Dissertation Goals

TIMSS expended significant effort and resources over a four-year period to develop the mathematics PSIs. Now that the PSIs have successfully been finalized and are part of an ongoing data collection effort in more than 30 countries around the world, it is the ideal time to look back at the development process and forward to reporting the results.

Given the new and pioneering effort represented by the PSIs and recognizing the author of this dissertations’ unique qualifications for carrying out this work, the Executive Directors of the TIMSS & PIRLS International Study Center were very supportive of the author conducting an in-depth investigation into the validity of the PSIs. As the TIMSS mathematics coordinator, the author was responsible for overseeing the development of the TIMSS 2019 mathematics assessments from start to finish, including the innovative PSI initiative. Thus, the author was well positioned to conduct this research on a timeline where the results could influence upcoming reporting and future development. This dissertation will not only contribute to decisions about how to analyze and report students’ achievement on the mathematics PSIs for TIMSS 2019, but also will

provide insights for future TIMSS assessments and other testing programs aspiring to capitalize on digital technology.

Overarching Research Question: Does adding the PSIs to the eTIMSS mathematics assessment enhance the validity of the TIMSS mathematics achievement scales at the fourth and eighth grades?

This question relates to planning future TIMSS assessments. If the answer is affirmative, then it would be important for TIMSS to consider adding PSIs as an integral part of future assessments. However, making that decision involves an investigation into the procedures and resources necessary to develop and administer valid PSIs. At this point, the development process was completed so the validity of test content could be thoroughly documented. Also, because a full-scale field test was conducted in 2018, the data collected in the field test could be used to investigate students' interactions with the assessment and explore the internal structure of the PSIs.¹ This latter area of investigation is especially pertinent in the context of reporting the PSI results and also has implications for future TIMSS mathematics assessments. The series of research questions to be examined within three major research areas are articulated below.

1) Validity Evidence Based on Test Content

- Did the methods used to develop the PSIs support a high-quality framework and coherent assessment instruments that minimize construct-irrelevant variance?
- Do the mathematics PSIs address the *TIMSS 2019 Mathematics Framework* and improve coverage of mathematics applying and reasoning skills?

¹ The fully documented TIMSS 2019 International Database will be available in February 2021 for further analyses related to these aspects of validity as well as achievement on the mathematics PSIs in relation to other variables and any consequences of having included the PSIs in TIMSS 2019.

2) Validity Evidence Based on Response Process

- Did the eTIMSS user interface, directions, and tools promote ease of navigation and consistency across the tasks?
- Are students' interactions with the eTIMSS mathematics instruments consistent with the cognitive processes the instruments were designed to elicit?
- Can the items that comprise the mathematics PSIs be scored reliably?

3) Validity Evidence Based on Internal Structure

- Do the properties of the mathematics items that comprise the PSIs differ from the regular eTIMSS mathematics items? And if so, how?
- How do the PSIs fit with the hypothesized factor structure underlying mathematics ability?

The research above will involve three key tasks: 1) examining and documenting the methods and procedures used to develop the PSIs, 2) investigating the characteristics of the PSIs in terms of content coverage and fidelity of student responses, and 3) using the eTIMSS field test data to learn more about the internal structure of the PSIs prior to TIMSS 2019 reporting. Another goal of this dissertation entails providing a summary of the lessons TIMSS learned in developing and implementing the mathematics PSIs and the contribution the PSIs have made to the field of digitally-based educational assessment.

Following this overview chapter, Chapter 2 provides a literature review of several topics relevant to this research. It begins with a description of the impact of large-scale assessments on mathematics education and assessment practices, providing the context for the TIMSS assessments and this dissertation. Next, it discusses the benefits of eAssessment, the challenges faced in designing digital, large-scale, assessments of mathematics problem solving and reasoning skills, and examples of other large-scale

assessments of problem solving skills. Then, it describes the test development framework that inspired TIMSS' approach to developing the PSIs and summarizes current best practices for gathering and evaluating validity evidence based on test content, response process, and internal structure that informed the methods used in this dissertation.

Chapter 3 documents the collaborative methods and procedures used to create and implement the PSIs for eTIMSS 2019, including the very ambitious field test conducted in March through May 2018. The data from the eTIMSS 2019 Field Test were used to address research questions about the response process and internal structure of the tasks. This chapter also describes the analysis methods used to further address the research questions.

Chapter 4 presents the results of the analysis methods described in Chapter 3, organized by research area. Finally, Chapter 5 provides a summary of the key findings and the implications for the eTIMSS 2019 mathematics PSIs. It then discusses the lessons learned throughout the development process and offers suggestions for capitalizing on digital technology in future assessments.

Chapter 2: Literature Review

This literature review consists of five main sections. The first section discusses the influence of large-scale assessments on mathematics instruction and assessment practices, demonstrating the global impact of TIMSS on mathematics education. The second section describes the status of the current digital revolution in educational assessment and summarizes the benefits of eAssessment, explaining why transitioning to eTIMSS was deemed worth the cost and effort for TIMSS. The third section describes the many challenges faced in designing the PSIs, including adhering to best practices in item writing, measuring problem solving and reasoning skills, developing problem contexts, and assessing mathematics via computers and tablets. The fourth section provides examples of other digital large-scale assessments of problem solving and inquiry that were considered in developing the eTIMSS PSIs. The final section describes the current best practices in test development and gathering and evaluating validity evidence that informed TIMSS' approach to developing the PSIs and the analysis methods used in this dissertation.

Mathematics Education and Large-Scale Assessment

Since the First International Mathematics Study (FIMS) in the 1950s, the mathematics education community has been a driving force behind large-scale international assessments. Around the world, mathematics educators agree that comparing both the inputs and outputs of education systems is essential to evaluating each system of interest because doing so can “reveal important relationships that would

otherwise escape detection within a single education system” (IEA, 2017). Further, international assessments such as TIMSS not only enable countries to view their own mathematics achievement in a global context, but also provide an opportunity for educators to learn about other countries’ approaches to mathematics education. Participating in international studies opens the door to sharing curricula, teaching methods, modes of assessment, and expectations for student achievement (Dossey, 2003; Kilpatrick, 1992; Robitaille, Beaton & Plomp, 2000).

Large-scale assessments also present a view of the mathematics content and processes that are most valued and serve as exemplars for measuring student achievement of these abilities (Suurtamm et al., 2016; Swan & Burkhardt, 2012). For example, in the *TIMSS 2015 Encyclopedia* (Mullis et al., 2016), many participating countries reported using the TIMSS frameworks, methods, and procedures in a variety of ways to reform their national systems. According to representatives from Chile, “TIMSS is regarded as a benchmark for assessment methodologies, evaluation frameworks, designing and coding of open ended questions, and results reporting, among other components of assessment” (Agencia de Calidad de la Educación, 2016). This perspective was echoed by many other participating countries, such as Armenia, Ireland, Kuwait, Morocco, and Serbia.

The content and design of the achievement items used in international large-scale assessments have a particularly strong influence on the implementation of curricula and classroom assessment. According to the International Congress on Mathematical Education (ICME), the nature and design of large-scale assessment tasks have “an enormous influence” on teachers’ instruction in terms of both the content and the types of

tasks that students experience (Suurtamm et al., 2016). The implemented curriculum will “inevitably be close to the tested curriculum,” making assessment a “uniquely powerful lever for forwarding large-scale improvement” (Barnes, Clarke & Stephens, 2010; Swan & Burkhardt, 2012). In the *TIMSS 2015 Encyclopedia* (Mullis et al., 2016), a number of countries including the Czech Republic, Italy, Jordan, Malaysia, Oman, and Slovenia, reported using released TIMSS items to inform the development of textbooks and train teachers in classroom assessment, providing clear evidence of this impact as well.

Given the significant and far-reaching influence of international large-scale assessments, it is imperative that the achievement items used in these studies are well aligned “not only with mathematics content, but also with mathematical processes and actions” that educators believe are most important (Swan & Burkhardt, 2012). However, despite assessment developers’ continued efforts to achieve this goal, there is a persistent dissonance between the knowledge and skills that are most valued by mathematics educators and the knowledge and skills that are assessed with large-scale assessments (Darling-Hammond et al., 2013; Swan & Burkhardt, 2012). In particular, the mathematics education community believes that the abilities to solve complex problems with multiple solutions and apply what is learned in the classroom to real-world scenarios are equally as important as any other content standards (Scherrer, 2015). Unfortunately, many large-scale assessments have the reputation of being predominantly comprised of straight forward multiple-choice items targeting lower-order skills such as computation and recall (Liljedahl, Santos-Trigo, Malaspina & Bruder, 2016; Scherrer, 2015).

This incongruity between what is valued and what is tested has detrimental effects on both teachers' assessment practices and students' achievement in and attitudes towards mathematics. When teachers see that the majority of the items on some large-scale tests are multiple-choice, algorithmic items, it suggests that these are ways mathematics should be taught and assessed in the classroom (Johansson, 2016). Students who mainly encounter routine items in the classroom have been found to have lower achievement than those who frequently engage in mathematics problem solving (Boaler & Staples, 2008; Hiebert & Wearne, 1993; Stein & Lane, 1996; Stigler & Hiebert, 2004). For example, in a study investigating the relationships between classroom assessment practices and achievement on the United States' national assessment (NAEP), Walcott, Hudson, Mohr and Essex (2015) found a strong, negative, correlation between the frequency with which teachers reported using multiple-choice items in the classroom and their students' achievement on NAEP. At the fourth grade, students of teachers who reported "never or hardly ever using multiple-choice assessments" in the classroom scored the equivalent of one grade level (11 points) higher than students of teachers who reported using multiple-choice items one to two times per week. At the eighth grade, this difference was equivalent to two grade levels (19 points).

To mitigate this unintended "backwash" of poorly designed tests, it is imperative that large-scale testing programs continue to strive to improve their achievement instruments (Swan & Burkhardt, 2012). Doing so will both enhance measurement of the abilities valued by educators and set a better example of best practices in assessment.

Benefits of eAssessment

Digital assessment offers new opportunities to increase the alignment between what is valued and what is tested. The bodies of literature surrounding test construction and other assessment programs' forays into eAssessment provide evidence of its potential to support superior measurement in a variety of ways. However, "we are still only at the cusp of realizing its full potential" (O'Leary et al., 2018). Bennett (1998, 2015) described three stages or "generations" as having occurred in the transition to digital assessment. The first generation was recreating paper-and-pencil item types in a digital environment to increase operational efficiency and build infrastructure; the second generation was introducing less traditional item formats and making initial attempts to measure new constructs; and the third generation is creating complex assessments including simulations and performance tasks that replicate the real world, allow natural interactions with digital devices, and assess skills in more sophisticated ways than ever before (Bennett, 1998, 2015). Fully achieving the third generation is predicted to take many years (Bennett, 2015; Pellegrino & Quellmaz, 2010; Redecker, 2013), but the available technology continues to become more complex and progress is being made.

Digital assessment can offer a multitude of benefits for large-scale testing programs that vary depending on the assessment goals and available technology. The following sections focus on those that are most relevant to eTIMSS and the mathematics PSIs—increased construct validity, student engagement, and operational efficiency.

Construct Validity

Multiple-choice items have long been disparaged as inadequate for measuring complex knowledge and skills (Archbald & Newmann, 1988; Birenbaum & Tatsuoka, 1987; Darling-Hammond & Lieberman, 1992; Lissetz & Hou, 2012; Measured Progress/ETS Collaborative, 2012), but continue to be prevalent in paper-based large-scale assessment because they offer the benefits of broad content coverage in a short amount of testing time, lower development and scoring costs, and reliable machine scoring (Jodoin, 2003; Measured Progress/ETS Collaborative, 2012). The alternative, constructed response item types, are hailed for providing better measurement, but are frugally used in paper-based large-scale assessments because they take more time for students to complete, are costly and time consuming to score, and are consequently less efficient than multiple-choice items (Bryant, 2017).

Digital delivery allows for many constructed response item types to be machine scored, which not only reduces costs associated with scoring, but also makes it possible to include more constructed response items in assessments, providing overall richer measurement (Dolan, Goodman, Strain-Seymour, Adams & Sethuraman, 2011; Lissetz & Hou, 2012; Measured Progress/ETS Collaborative, 2012; Scalise & Gifford, 2006). Constructed response items permit a range of answers, requiring students to organize ideas rather than recognize them. Therefore, they offer greater insight into how students approach problems and allow for partial credit scoring and collection of more diagnostic information (Lissetz & Hou, 2012). Further, these less constrained item formats are typically more effective in differentiating between students with higher and lower

achievement (i.e., more highly discriminating) than multiple-choice items. For example, in TIMSS 2015, the one-point mathematics constructed response items had an international average discrimination of 0.46 at the fourth grade and 0.50 at the eighth grade, while the multiple-choice items had an international average discrimination of 0.39 and 0.41, respectively, at the fourth and eighth grades (Foy, 2017).

Research suggests that technology-enhanced assessment also has the potential to improve construct validity by way of innovative item types, such as drag and drop, sorting, or multiple-selection, as well as on-screen tools such as rulers and calculators that better support construct representation (Huff & Sireci, 2001; Parshall, Harnes, Davey & Pashley, 2010; Sireci & Zenisky, 2006). In particular, well-designed technology-enhanced items have been found to improve measurement of higher-level skills including reasoning, synthesis, and evaluation, especially if they involve a real-world context. Such items can be used to elicit active construction of knowledge by requiring direct interaction with stimuli, and consequently tap different cognitive constructs than traditional multiple-choice items and reduce the effect of guessing (Dolan et al., 2011; Huff & Sireci, 2001; Measured Progress/ETS Collaborative, 2012; Strain-Seymore, Way & Dolan, 2009). For example, in a comparison between the Item Response Theory (IRT) information provided by innovative item types (i.e., multiple-selection, drop and connect, and create-a-tree) and multiple-choice items, Jodoin (2003) found that the innovative item types provided considerably more expected IRT information across all ability levels. The mean expected information for innovative items

was 0.32, compared to 0.17 for multiple-choice items, demonstrating the potential for innovative item types to deliver more precise measurement (Jodoin, 2003).

Administering assessments on digital devices also allows for efficient capture of additional information beyond responses to items, referred to as “event data” (e.g., time on task, series of clicks, use of tools). These data can provide new insights into students’ abilities and interactions with assessments that were not feasible or possible to collect in paper-based assessment (Greiff, Niepel, Scherrer & Martin, 2016; Shu et al., 2017).

Student Engagement

Developing items for computer- or tablet-based testing also makes it possible to incorporate colorful graphics, interactive features and tools, videos, and animations into the item stimulus and response space. These features can make assessments more engaging for students because they are more hands-on, visually appealing, and authentic than paper-based booklets (Bryant, 2017; Dolan et al., 2011; Measured Progress/ETS Collaborative, 2012; Parshall et al., 2010; Strain-Seymour et al., 2009). They can also help to increase construct validity by allowing for questions to be asked that cannot be posed on paper, such as items about short videos or simulations.

Adding attractive and interactive features to low-stakes assessments is particularly important because these features can help counter persistent issues with student motivation and effort that commonly arise in large-scale assessments and present a threat to the reliability and validity of test scores (Wise, Pastor & Kong, 2009). Research on response time and effort indicates that examinees who exhibit “rapid guessing behavior” are more likely to engage with items that have attractive surface features and minimal

text, both of which are common in digital items, as they are more appealing in a quick appraisal of the task (Wise et al., 2009).

Operational Efficiency

Once the challenging task of developing the software infrastructure to conduct a digital assessment is complete, the first benefit of transitioning to digital assessment is typically increased efficiency in creating and carrying out the assessment (Bennett, 2015). In a coherent system, tasks such as creating items, assembling instruments, collecting data, assigning scores to machine-scored items, distributing responses to scorers for human-scored items, performing quality control, and monitoring test security can be completed in a more effective manner than is possible with paper-based testing (Bennett, 2015; Bryant, 2017). When these tasks require less effort, developers can focus more of their attention on the test content. Digital delivery also eliminates the costs of printing, shipping, and collecting test booklets, which are considerable expenses in large-scale assessment (Quellmalz & Pellegrino, 2009).

In sum, computer-based testing offers the potential to simultaneously leverage the machine-scoring benefits of multiple-choice items and measurement benefits of constructed response items to ameliorate design constraints that have historically been a limiting factor in paper-based large-scale assessment. eAssessment aids assessment programs in addressing criticisms that large-scale assessments “cannot and do not reflect the breadth and depth of knowledge, skills, and abilities associated with a construct of interest” (Jodoin, 2003, p. 1) and increasing student engagement. For TIMSS, transitioning to computer-based testing also increased efficiency in translation and

instrument verification, delivery, and data entry. Using the IEA’s eAssessment system, these processes for eTIMSS 2019 were integrated into a single system, bringing together many aspects of TIMSS operations.

Challenges in Developing the eTIMSS Mathematics PSIs

Developing the infrastructure to shift from paper-and-pencil to computer- and tablet-based testing is a huge undertaking (Bennett, 2015; Drasgow, Luecht, & Bennett, 2006). Striving to create cohesive sets of mathematics items that fulfill the ambitious aspirations for the PSIs in a developing system presented even greater challenges beyond the baseline requirements for designing valid and reliable assessment items. In addition to covering the mathematics content in the framework and adhering to best practices in item writing, each PSI also needed to measure problem solving and reasoning skills, be situated within a context that is engaging and appropriate for students around the world, and leverage technology. Developing the eTIMSS mathematics PSIs therefore necessitated additional attention to assessing higher-order skills, developing contexts that are appropriate for an international audience, and assessing mathematics on a computer.

Adhering to Best Practices in Item Writing

Developing valid, reliable, and unbiased items to measure mathematics ability in an international context is a complicated undertaking, regardless of the mode of administration. As explained in the *TIMSS 2019 Item Writing Guidelines* (Mullis, Martin, Cotter & Centurino, 2017), writing good items “requires imagination and creativity, but at the same time demands considerable discipline in working within the assessment frameworks and following guidelines for item design” (p. 3). First and foremost, it is

essential to make certain that each assessment item can be obviously related back to a content topic and cognitive process as described in the assessment framework, has appropriate difficulty for the target population, and can be reliably scored (by human or machine). It is also imperative that an item makes clear what is being asked of the respondent, is feasible to complete in a reasonable amount of time, uses grade-appropriate language, and avoids cultural, gender, or geographical bias (Mullis et al., 2017). For selected-response item types, there must be only one correct answer key (one of the options for traditional multiple-choice, more than one for multiple-selection items), the incorrect answer options or “distracters” must be realistic alternatives to the key(s), and all options must be phrased or depicted in a consistent format to avoid making any option stand out from the rest. For constructed response item types, it is imperative to design scoring guides with clear distinctions among correct, incorrect, and if applicable, partially correct answers, which human scorers or machines will be able to apply to student responses with a high degree of reliability (Mullis et al., 2017).

Assessing Problem Solving in Mathematics

The complexity of posing a problem to be solved and delineating its solution go hand in hand—the more complex a problem is intended to be, the more difficult it is to develop (Crespo & Sinclair, 2003). Knowledge-based mathematics items are straightforward, presented without a context, and ask students to recall, recognize, classify, compute, or retrieve information (Lindquist et al., 2017). These items are typically the simplest to pose as well as the simplest to answer, as they are factual (Crespo & Sinclair, 2008) and often involve minimal text (e.g., “naked” computation or

algebra problems). Application items ask students to make connections between their knowledge and the situation at hand or involve multiple steps to a solution (Lindquist et al., 2017). Developing application items necessitates creating a context or an additional layer of complexity to require more than one step, resulting in some additional effort in item writing. Still, application items are designed with the intention of maintaining a clear connection between the problem and process to a solution, so they can be relatively straightforward to develop as well.

A question is considered a “problem” when it lacks a readily available or routine method to a solution (Greiff, Holt & Funke, 2013; Mayer & Alexander, 2016). Building from this definition, “problem solving” can be defined as “the process of transforming the given state into the desired goal state” (Lovett, 2002), which involves two steps: 1) establishing a representation of the problem (knowledge acquisition), and 2) implementing a solution process (knowledge application) (Greiff et al., 2013; Klahr & Dunbar, 1988; Novick & Bassok, 2005).

Historically, developing questions to measure students’ ability to integrate, synthesize, and creatively apply knowledge to novel situations outside the classroom has proven to be difficult because traditional problem solving tasks do not easily translate into valid and reliable achievement items (Bennett, Persky, Weiss & Jenkins, 2003; Greiff et al., 2013). In an assessment situation, it is fundamental to collect measures of student ability and thus large amounts of response time cannot be devoted to letting students play with interactive features of the task and give up. Classroom-based problem solving tasks are often time consuming, challenging to score, and rely on teacher led

instruction. As such, they generally measure multiple intertwined skills rather than a single construct (i.e., are multidimensional) and trying to modify such tasks into assessment items often violates the statistical assumptions underlying the models used in analyzing student responses (i.e., that the items are independent and unidimensional) (Bennett et al., 2003).

Test developers must carefully balance tradeoffs in introducing a degree of structure into a problem solving task (i.e., scaffolding toward the answer) and maintaining authenticity by keeping structure to a minimum. Consequently, designing assessment items that elicit logical and systematic thinking requires diligent attention to the parameters of the scenario, the information that is given and withheld, and the language used to convey this information to the solver. Taken together, these requirements make developing successful achievement items to measure problem solving skills substantially more difficult than developing items to measure knowing and applying skills.

Developing Problem Contexts

Research on best practices in mathematics assessment suggests that relevant, age appropriate, problem situations or contexts can increase student engagement, motivation, and perseverance, which is particularly important in low-stakes testing scenarios because the results will have greater validity when students give their maximum effort (Bennett, 2014; Greaux, 2013; Nijlen & Janssen, 2015; Sugrue, 1995; Wise et al., 2009). However, for an item in context to be considered valid, the context must be equally familiar to all examinees (Sugrue, 1995), which is a substantial challenge for TIMSS

because the assessments are administered in over 60 countries. Presenting students with items situated within unfamiliar contexts has been shown to introduce construct-irrelevant variance and bias, which can result in different response patterns among subgroups of students with the same achievement score, or differential item functioning (DIF) (Boaler, 1993; Greairex, 2013).

The challenge of developing problem contexts is further complicated when designing a single context to span a series of items. The context must not only be complex enough to require multiple questions, but also not overly complicated such that it unnecessarily increases reading demand or deters examinees (Cormier, Yeo, Christ, Offrey & Pratt, 2016; Sugrue, 1995). Consequently, developing problem solving contexts was one of the first challenges in developing each mathematics PSI and the contexts for the tasks continued to evolve throughout the development process.

Assessing Mathematics on a Computer

Despite the benefits of innovative item types, some traditional paper-based mathematics item formats become more cumbersome for students to complete on a digital device. For example, research has shown that items requiring students to draw, enter lengthy amounts of text, and type equations or formulas are more difficult on a computer or tablet than on paper (Sandene, Bennett, Braswell & Oranje, 2005). Further, items that necessitate “scratch work” may be more difficult on a computer/tablet because students need to transfer information from the screen onto scratch paper and then the answer back to the digital device, increasing the risk of transcription errors (Russell, Goldberg & O’Conner, 2003).

In other cases, features unique to computer-based items have been shown to reduce item difficulty. In a small-scale comparison of student achievement on technology-enhanced mathematics items and their paper-based equivalents, Therelfall, Pool, Homer and Swinnerton (2007) found that students performed better on several types of technology-enhanced items, such as those involving draggable number cards to order numbers or complete number sentences. With equivalent mathematics content across modes, the authors hypothesized that these items were less difficult on the computer because being able to drag the cards reduced the amount of information students needed to hold in their working memory and made exploring possible solutions more accessible by eliminating the need to cross out or erase (Therelfall et al., 2007).

The body of literature surrounding enhanced item types is growing, but there is still a lack of certainty about how test takers interpret and interact with novel item types and their resulting measurement quality (Parshall & Becker, 2015). The Partnership for Assessment of Readiness for College and Career assessment consortium (PARCC) estimated that development costs for enhanced item types are two to five times greater than those of traditional multiple-choice items, making it particularly important to carefully consider the utility of each enhanced item for measuring the target construct (Russell, 2016). With limited information on best practices for newer item types, developing enhanced items that are valid, reliable, and suitable for an international assessment was a substantial challenge for eTIMSS 2019 and extremely resource intensive.

Examples of Large-Scale Problem Solving Assessments

In response to the growing importance of critical thinking, information and communication in technology, collaboration, and problem solving in today's workforce, NAEP (National Assessment of Educational Progress) and PISA (Programme for International Student Assessment) began to assess these 21st century skills with computer-based complex problem solving (CPS) tasks in 2003 and 2012, respectively. In CPS tasks students are provided with a set of tools that they may choose how to apply in a flexible environment to solve a problem. This minimally constrained design aims to simulate a real-world problem scenario by making it possible to approach the problem in a variety of ways (Herde, Wüstenberg & Greiff, 2016). Because CPS tasks typically only ask for one or two direct answers, student achievement is primarily measured via log files of students' observable actions (e.g., series of clicks) referred to as "event data" en route to the solution (Organisation for Economic Co-operation and Development [OECD], 2017).

CPS tasks fundamentally differ from the PSIs in the constructs they are designed to measure as well as in the way they are scored. However, NAEP and PISA's CPS assessments are currently the most relevant examples of digital large-scale assessments of higher-order thinking skills, so the benefits and disadvantages of these innovative forms of assessments were considered in developing the eTIMSS PSIs.

NAEP – Problem Solving in Technology-Rich Environments

NAEP began its venture into digital assessment of problem solving in 2003 as a part of an experimental technology-based assessment project. This portion of the

assessment was designed to measure “Problem Solving in Technology-Rich Environments” (TRE), which was conceptualized as “the intersection of content areas and technology environments” (Bennett, Persky, Weiss & Jenkins, 2007). The assessment consisted of one problem scenario, using helium balloons to explore outer space, with two variants—a search task, in which students look up information to answer a question about the use of these balloons, and a simulation task, in which students design, run, and interpret the results of an experiment with the balloons (Bennett et al., 2007). NAEP described these tasks as “partial inquiry” because they imposed some constraints on students’ actions for the purposes of limiting testing time and safeguarding against uninterpretable data (Bennett et al., 2003), but the tasks were relatively flexible, as students worked on a single screen for the entirety of the testing session with limited scaffolding to direct their approach. Students were scored based on their observable actions within the problem environment (e.g., use of search terms and tools) as well as their responses to a short series of “motivating problems” (i.e., traditional multiple-choice items embedded in the task), which were added to increase the likelihood that the tasks would provide adequate measurement of students’ scientific inquiry skills (Bennett et al., 2007).

PISA – Complex and Collaborative Problem Solving

PISA introduced “Complex Problem Solving” as a minor assessment domain in 2012, then added “Collaborative Problem Solving” as separate domain in 2015 to improve coverage of the 21st century skills the problem solving assessment was designed to measure. The format of PISA’s Complex Problem Solving tasks in 2012 was similar to

NAEP's TRE tasks, but with a broader variety of item formats and interactive features. For example, in the task "Climate Control," students determined how the controls on an air conditioner work by experimenting with buttons and viewing the impact on the temperature and humidity, then created a diagram to explain the functions of the buttons (OECD, 2017). Each Complex Problem Solving task included one or two direct questions, and all other measures of student achievement were extracted from log files.

In Collaborative Problem Solving tasks, students work with a computer agents via a chat box and shared workspace to solve a problem. They are evaluated based on their interactions, the extent to which they establish and maintain shared understanding and team organization throughout the process, and the appropriateness of the actions taken to solve the problem (OECD, 2017). For example, in a sample PISA 2015 Collaborative Problem Solving scenario, students collaborated with a computer agent to find the optimal conditions for an aquarium environment. The information needed to solve the problem was divided between the student and computer agent, such that it was necessary for the student to work with the computer agent to answer the questions (OECD, 2017).

Benefits and Challenges of Complex Problem Solving Tasks

Minimally constrained and highly realistic CPS tasks offer the benefit of being very authentic, engaging, and capable of measuring complex constructs that are not feasible to measure in large-scale assessment with traditional item formats (Greiff et al., 2016). CPS tasks can produce immense amounts of data by capturing students' every step in solving a problem, offering an entirely new source of information that has potential to increase the validity of test scores and provide deeper insight into students' cognitive

processes (Herde et al., 2016; Shu, Bergner, Zhu, Hao & von Davier, 2017). For example, event data may help to illuminate how test performance evolves or how differences in countries' performance on such tasks are grounded in behavioral differences that may relate to educational policy (Herde et al., 2016).

However, highly unstructured CPS tasks have many of the same issues as classroom-based problem solving tasks. CPS tasks are time consuming to complete and challenging to score, reducing the number of items a student can feasibly take and consequently weakening the discriminating power of the assessment and diminishing the reliability of scores (Funke, 2009; Greiff, Wustenburg & Funke, 2012; Herde et al., 2016). Also, in cases where a single misstep impacts a students' trajectory through the task, students' scores may be impaired as it becomes difficult to demonstrate any skills after an initial mistake (Fischer et al., 2015; Greiff et al., 2012; Herde et al., 2016). These issues can result in unintentional local dependence and multidimensionality, which compromise the validity of test scores (Bennett et al., 2003).

Further, despite the high expectations for event data, little progress has been made thus far in extracting useful information from the complex log files (Greiff et al., 2016; Shu et al., 2017). Currently, a variety of methods for analyzing and interpreting log files are being explored, but most are still largely experimental and require strong assumptions, undercutting the utility of such data for generating achievement scores (Shu et al., 2017).

“Beyond Constructed Response Items”

Another class of items has been identified in the literature for having promising potential to strike a beneficial balance between the structure afforded by standalone achievement items and the increased construct representation offered by CPS tasks (Huff & Sireci, 2001; Parshall et al., 2010; Sireci & Zenisky, 2006). According to Parshall et al. (2010), a “beyond constructed response” item set is a series of items presented together within the structure of a single context. These item sets are necessarily less authentic than “real world problems,” but they offer a more reliable assessment approach that may be well-suited for domain-specific problem solving skills. Beyond constructed response item sets can include a variety of item types with varying degrees of structure, allowing for in-depth investigation of a problem scenario with a reduced risk of uninterpretable data (Parshall et al., 2010). The mathematics PSIs fit well with this classification, although the tasks were not designed with this specific label in mind.

Still, beyond constructed response item sets present their own design constraints and require considerably more development effort to achieve valid measurement than discrete assessment items (Parshall et al., 2010). For example, when TIMSS experimented with a series of paper-based extended problem solving and inquiry tasks in 2003, the TIMSS & PIRLS International Study Center reported that developing suitable problem contexts that required sustained study and challenged students, but were not overly intimidating as to discourage students from engaging with the task, was a difficult balance to achieve (IEA, 2005). Further, maintaining independence among the items to

avoid issues in analysis while also providing scaffolding and adhering to the problem context presented another formidable challenge (IEA, 2005).

In 2014, NAEP designed a series of “Technology and Engineering Literacy” (TEL) tasks that may also be considered an example of beyond constructed response. Unlike the earlier TRE tasks, each TEL task was comprised of multiple screens through which students progressed toward a solution. These tasks included more features to keep students on track (e.g., pop-up notifications after a period of inactivity, requirements to answer a question before moving on to the next) and the item pool included a range of short (10 minute), medium (20 minute) and long (30 minute) tasks (NAEP, 2014a; 2014b). Although this format was less authentic than the first TRE assessment, these features proved to be successful in acting as a stronger safeguard against inexplicable data.

Summary of Benefits and Challenges of eAssessment

Meeting the aspirational development goals for the eTIMSS 2019 mathematics PSIs was a substantial undertaking. Developing suitable problem contexts for an international audience, creating the series of items to guide students through the problem scenario, and determining how to capitalize on technology to support good measurement all presented challenges along the way. However, the eTIMSS PSIs were expected to serve the important purpose of increasing coverage of complex areas of the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017) beyond what is possible with traditional achievement items, making the development cost worth the effort.

Promoting and Demonstrating Assessment Validity

Ensuring that the mathematics PSIs provide valid measurement of mathematics achievement required simultaneous attention to a variety of design features and constraints, which TIMSS managed by keeping in mind the salient aspects of the evidence-centered design (ECD) framework (Mislevy, Almond & Lukas, 2003). Adhering to the ECD framework is a labor intensive and time consuming process (Huff, Steinberg & Matts, 2010) that TIMSS cannot feasibly abide by in every assessment cycle. However, the TIMSS development process is inspired by the principles of ECD and includes many of the recommended steps for establishing validity by design.

Under the ECD framework, assessment is viewed as a form of evidentiary reasoning, in which each target measurement is articulated as a claim to be made about the student (Huff, Steinberg & Matts, 2010; Mislevy et al., 2003). The framework provides the structure for an iterative development process that guides test developers in certifying that “the way evidence is gathered and interpreted is consistent with the underlying knowledge and purposes the assessment is intended to address” (Mislevy et al., 2003, p. 2). Using this approach aids test developers in capitalizing on current advances in student learning and assessment, formulating design specifications, framing the item writing process, and coordinating participation in development (Huff et al., 2010; Mislevy et al., 2003).

The ECD development process begins with identifying and prioritizing the content and skills that comprise the target construct the assessment is intended to measure (*domain analysis*) and delineating reasonable and observable forms of evidence that can

be collected via assessment items to support the target claims (*domain modeling*) (Huff et al., 2010). Next, a *Conceptual Assessment Framework* (CAF) is established, which defines and connects all parts of the assessment (Messick, 1994; Mislevy et al., 2003).

The CAF includes five key parts, referred to as principle design objects:

- The *Student Model* defines one or more unobservable variables related to the knowledge, skills, and abilities (KSAs) the assessment is intended to measure and how student achievement of the KSAs will be expressed (Mislevy & Riconscente, 2005).
- The *Evidence Model* describes how to illicit evidence from students' work in the context of the assessment, rules for scoring the work, and how each piece of information directly characterizes an aspect of performance and conveys information about the target claim (Mislevy & Riconscente, 2005).
- The *Task Model* serves as a template for the items, including specifications for stimulus material, the work students will be asked to produce, the assessment conditions, and presentation (Mislevy et al., 2003).
- The *Assembly Model* describes the combination of items that comprise forms of the assessment, which can include a variety of item characteristics (e.g., content, cognitive demand, format) (Huff et al., 2010; Mislevy et al., 2003).
- The *Presentation Model* describes how the tasks will appear, including the mode of delivery, specifications for the delivery platform, the tools provided to test takers, and the timing for the testing sessions (Mislevy et al., 2003).

The *Delivery Model* encompasses all five of the principle design objects and describes issues that cut across the CAF, including administrative constraints, security procedures, and data recovery protocols (Mislevy et al., 2003).

Once an assessment is developed, validity evidence must be gathered to evaluate the extent to which the proposed interpretations of scores on an assessment are valid (AERA, APA & NCME, 2014). The following sections describe current best practices for collecting evidence for the types of validity that were evaluated in this dissertation—test content, response process, and internal structure.

Validity Evidence Based on Test Content

Evidence based on test content is gathered through logical or empirical analysis of the accuracy with which the test content, including item formats, themes, wording, directions, and scoring, represents the target construct (AERA, APA & NCME, 2014). In essence, this strand of evidence is an evaluation of the authenticity and coherence of the assessment framework and specifications. Content validity can be promoted by taking into account the ECD framework. In particular, it is essential to clearly define the target construct, specify the items needed to measure it, and establish guidelines for item writers, designers, and programmers to support the development of high quality instruments and minimize construct-irrelevant variance (Dolan et al., 2011; Dolan, Strain-Seymour, Way & Rose, 2013). Evidence of content validity can be provided by thoroughly documenting the methods used to certify that the assessment content meets these criteria, including through the use of assessment blueprints, user interface

specifications, item writing guidelines, expert reviews of test content, test administration manuals, scoring guides, and scorer training.

An assessment blueprint gives a detailed outline for the composition of the test in terms of the percentage of score points allocated to each topic area in the assessment framework, cognitive skill, and item type, and can be used to demonstrate the connection between the test content and the assessment framework (AERA, APA & NCME, 2014). Development of an assessment blueprint should be informed by relevant curriculum standards and instructional approaches. Once established, every item should be classified on all dimensions in the blueprint (e.g., content topic, cognitive demand, format) to ensure that each item addresses knowledge and skills in the assessment framework, and that altogether, the group of items cover the target construct as planned (Dolan et al., 2013).

Item writing guidelines describing the available item formats, desirable item characteristics, and issues to avoid in item writing (e.g., context-specific vocabulary, unnecessary graphics, vague wording) should be established to support item writers in developing high quality assessment items to cover the assessment blueprint. For digital assessments, item writers, designers, and programmers should all be provided with a user interface template so that screen “real estate,” layout, and aesthetics may also be taken into account in designing items (Dolan et al., 2013).

Once an initial item pool is developed, subject matter experts should evaluate the items in terms of content and cognitive demands to determine whether each item appropriately samples the target construct as described in the assessment framework and

avoids the inclusion of irrelevant features that could interfere with measurement of the target construct (AERA, APA & NCME, 2014). Dolan et al. (2013) identified several positive characteristics of suitable test content that experts should consider in their review—relevant, representative, realistic, synergistic, clear and unambiguous, free of bias, and appropriate time and task load. Expert review may also involve comparing the items being developed and items designed to measure the same construct to determine the extent to which the test content is consistent with existing assessments (AERA, APA & NCME, 2014; Cook, Zendejas, Hamstra, Hatala & Brydges, 2014).

To ensure that the items in the final instrument possess desirable measurement properties (i.e., appropriate difficulty and discrimination, are unbiased), all items should be field tested prior to final selection (Parshall & Becker, 2015; Wilson, 2005).

Validity Evidence Based on Response Process

Generating evidence based on response process requires theoretical and empirical analysis of the relationship between the expected actions of test takers, administrators, and scorers in carrying out the test and how these parties interact with the test in practice (AERA, APA & NCME, 2014). Particularly when introducing new and potentially unfamiliar modes of assessment and item types, it is critical to evaluate the extent to which the interactions between all relevant parties and the test are operationalized as expected to ensure that interpretations of test scores are valid (Kreiter, 2015; Massachusetts Department of Elementary and Secondary Education [DESE], 2013, 2016). The body of literature surrounding best practices in gathering evidence of response process validity is still relatively limited, as it is the newest addition to *The*

Standards, but suggests several viable strategies for addressing this strand of evidence (Cizek, Rosenberg & Koons, 2008; Padilla & Benítez, 2014).

When developing a digital assessment, it is essential to establish user interface specifications to promote consistency across the tasks students will engage with to reduce cognitive demands associated with determining how to interact with the assessment (Dolan et al., 2013). Specifications should include a uniform layout for the interface design (e.g., standard location for navigation buttons, standard font) and specifications for the appearance and functionality of interactive components, including available tools appropriate to the tasks (Dolan, et al., 2013). It is important for designers to adhere to current best practices for universal design in establishing these specifications to promote accessibility for all test takers and incorporate system-level accessibility features when possible (Dolan et al., 2013).

For test takers, the response process is comprised of how students think through, interpret, and respond to items, and the degree to which students' problem solving strategies are consistent with those envisioned by the test developers (Desimone & Carlson, 2004; Gorin, 2006; Hopfenbeck & Maul, 2011; Kane, 2006). These interactions are not only influenced by the test content, but also by students' familiarity with and usability of the user interface, as well as the clarity of directions, which are especially important in technology-enhanced assessment to minimize the unintended impact of computer familiarity on test scores (Auewarakul, Downing, Jaturatamrong & Praditsuwan, 2005; Dolan et al., 2011).

Cognitive interviews or focus groups with examinees are the most direct methods of gaining insight into how test takers interact with a test in practice (Dolan et al., 2011; Hopfenbeck & Maul, 2011). Cognitive laboratories can be particularly useful in evaluating the usability of computer interface for enhanced item types by including questions such as “What features of the item made it easy to use or difficult to use?,” “How does this item compare to items that you typically see on a test?,” and “Which item would you rather answer—this one or a multiple-choice item? Why?” (Dolan et al. 2011). Systematic observations of testing sessions also can be used to gather evidence of students’ interactions with the test content and user interface, providing insight into confusing or frustrating features of the test and student engagement (Smarter Balanced Assessment Consortium, 2016). Pilot testing provides an opportunity to test the operations associated with the assessment, serving as a “dress rehearsal” for all systems involved in test delivery and administration (Mullis, Cotter, Fishbein & Centurino, 2016; Parshall & Becker, 2015).

The amount of time test takers spend on each item also can be used to evaluate whether test developers’ hypotheses about the cognitive complexity of items are consistent with the time needed to complete the items in practice (Cepeda, Blackwell & Munakata, 2013; Padilla & Benítez, 2014). Digital assessment allows for screen by screen timing data to be captured, enabling closer inspection of this relationship. For example, in a validation study of the Massachusetts Adult Proficiency Test (MAPT), Wang and Sireci (2013) used timing data to identify a relationship between the expected

complexity of the cognitive operations involved with the items and examinees' response time, which was mediated by item difficulties (Padilla & Benítez, 2014).

For low-stakes assessments such as eTIMSS, time on task may also be used as a proxy for student effort, which is an essential prerequisite for valid and reliable measurement of student achievement (Kupiainen, Vainikainen, Marjanen & Hautamäki, 2014; Lee & Chen, 2011). If scores are to be interpreted as what students know and can do, it is critical that the responses provided during testing sessions are an accurate representation of student ability. Kupiainen et al. (2014) demonstrated the potential impact of motivation through a study of time on task in a low-stakes assessment for ninth grade students in Finland. Taking into account prior student achievement (GPA), the authors found that time on task accounted for 20 percent of the total variance in students' test scores, and mediated the effects of GPA and self-reported negative attitudes toward the test on students' test scores (Kupiainen et al., 2014).

For test administrators, facilitating response process validity involves following test administration protocols and promoting test security to uphold the integrity of the data collected (Cook et al., 2014). Therefore, it is critical that detailed test administration manuals and scoring materials are developed to support test administrators in delivering the assessment as intended. Observations of the testing sessions can be useful in gathering evidence of the test administrators' behaviors, which may increase understanding of the appropriateness of the test administration protocols and adequacy of measures used to ensure test security (Auewarakul, et al., 2005; Cook et al., 2014).

The response process validity of an assessment also depends on the reliability and validity of the scores assigned to responses to the items on the test. eAssessment allows for a wider variety of constructed response item types to be machine-scored, which can increase the quality of assessment data by largely eliminating the inevitable inconsistencies that arise in human scoring (Yamamoto, He, Shin & von Davier, 2017). However, developing machine scoring specifications requires meticulous attention to detail in defining the range of possible responses for each score code and confirming that all responses are assigned the appropriate score. Methods of validating machine scoring rules depend heavily on the item format, but in general, technical reports of other large-scale testing programs (e.g., PISA) highlight the importance of testing machine scoring systems prior to data collection and having more than one individual or group apply the scoring rules and comparing the results to check for agreement (Yamamoto et al., 2017).

For human-scored items, response process validity can be supported by developing high quality scoring guides, providing training on how to apply the guides, requiring quality control throughout the scoring process, and using inter-rater reliability analysis to evaluate the degree of agreement among independent scorers (Auewarakul et al., 2005; Cook et al., 2014; Mullis et al., 2016). Applying scoring guides to student responses is a subjective task, and humans are susceptible to fatigue, error, or opinions that can result in more lenient or severe applications of scoring guides (Yamamoto et al., 2017). To promote reliable scoring of constructed response items, focused scoring guides that explicitly match the criteria delineated in the assessment framework should be developed and appropriately qualified scorers (e.g., teachers familiar with the subject

matter on the test) should be trained to understand the procedures and general scoring principles necessary to accurately apply the scoring guides (Kuo, Wu, Jen & Hsu, 2015; NAEP, 2017).

Validity Evidence Based on Internal Structure

Validity evidence based on internal structure is primarily obtained through analysis of the interrelationships among the items on the test and between the items and the target construct to determine the extent to which the observed relationships match the hypothesized structure of the construct (AERA, APA & NCME, 2014). This form of evidence is considered to be of paramount importance in upholding the validity of test scores and should be evaluated using a variety of the available techniques (Wilson, 2005).

Measurement Properties of Items

First, it is important to assess the measurement properties of the individual items. The difficulty and discrimination of each item should be evaluated to determine whether it is appropriately difficult for the target population and whether it is successful in differentiating between high and low performing students (AERA, APA & NCME, 2014; DESE, 2013, 2016; Cook et al., 2014; Evers, Sijtsma, Lucassen & Meijer, 2010). The item difficulties should also be considered as a group, to ensure a varying spread across the range of abilities in the target population (AERA, APA & NCME, 2014). Differential Item Functioning (DIF) analysis should be used to identify items on which sub-groups of students with similar overall scores perform substantially different to identify any items that may be biased against particular sub-groups, as scores on an assessment with bias items cannot be considered valid (AERA, APA & NCME, 2014; DESE, 2013, 2016).

Underlying Factor Structure

Second, it is critical to evaluate the relationships among the items that comprise the assessment (AERA, APA & NCME, 2014; Bennett, Persky, Weiss & Jenkins, 2010; Cook et al., 2014; Evers et al., 2010; Kind, 2013; Kuo et al., 2015). This can be done using a variety of different measurement models that fall under the overarching framework of generalized linear and latent mixed models, which includes both factor models and item response models (de Ayala, 2009; Reise, 2012; Toland, Sulis, Giambona, Porcu & Campbell, 2017). This framework assumes that an unobservable trait, or latent variable, exists and can be measured through responses to items, which are regarded as observable manifestations of the trait. It also assumes that items are an imprecise measurement tool, so there is always some error associated with the observed responses.

The specific modeling approach for an assessment should be selected based on *a priori* theory about the items and target constructs, the scale of the observed item response data (e.g., continuous, binary, ordinal), and other characteristics of the data such as the number of items, responses, and extent of missing data. The following sections describe several of the most commonly used techniques that were considered for this dissertation.

Exploratory Factor Analysis

Exploratory factor analysis (EFA) is multivariate technique that uses the correlation or covariance matrix of item responses to model the common variance among the items and the unobservable latent variables, or factors, that the items are designed to

measure. This data-driven approach is intended to help determine the number of factors that influence responses to the items on an instrument when there is no strong *a priori* theory about these relationships, but some conceptual justification for analyzing the group of items together (Hair, Black, Babin & Anderson, 2014). A variety of techniques may be used to extract the factors from the data (e.g., principal axis factoring, weighted least squares, maximum likelihood estimation) which will initially produce the same number of factors as items in the model. After the factors have been extracted, it is up to the researcher to determine the number of factors to retain, which may be decided based on criteria such as the interpretability of the factor structure (i.e., best conceptual structure) or the total amount of variance explained (i.e., practicality of the solution) (Hair et al., 2014).

Traditionally, EFA is used to help establish a theoretical basis for a confirmatory model, but when there is already a strong theory about the items and constructs (e.g., the TIMSS mathematics items and mathematics ability) it is not necessary to use EFA first (Hair et al., 2014).

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA; Jöreskog, 1969; 1971a) should be used to evaluate the fit of a hypothesized factor structure for the observed responses to items when there is a strong *a priori* theory about the structure of the assessment based on prior knowledge or exploratory analysis (Brown, 2014). In CFA, the researcher specifies the expected relationships among the items and latent variables before fitting the model by assigning each item to one of a number of factors in the model that the item is expected to

measure. Estimating the model allows for testing the fit of the data to the theory, which can be evaluated based on a variety of fit indices that take into account elements of the model that may impact model fit (e.g., sample size, complexity of the model) and the magnitude of the factor loadings, which indicate the strength of the relationship between each item and the factor to which it was assigned.

CFA allows for a more parsimonious solution to be tested than is possible with EFA because the number of factors, item-factor relationships, and error covariances are all pre-specified (Brown, 2014). The CFA model can also account for correlations among the factors, which are common when measuring multiple facets of a single construct. Therefore, when there is a strong hypothesis about the underlying structure of the data, CFA is an optimal technique for determining the dimensionality of a scale or group of sub-scales (Brown, 2014).

However, under the traditional CFA model the responses to each item can only be attributed to a single latent variable, which is often an unrealistic condition for educational assessments. Most educational and psychological assessments are inherently multidimensional due to either item multidimensionality (Reckase, 2009) or the intended content or construct structure of the assessment (Ackerman, Gierl & Walker, 2003), and therefore require more complex models that are consistent with their underlying structure (de la Torre & Song, 2009). Also, the traditional CFA model is not intended to be used with categorical observed variables, which are common on achievement tests. Categorical data have a restricted range of possible values (e.g., 0 and 1) and therefore applying a linear factor model to categorical observed variables will result in implausible estimates

for factor scores as well as violations of the assumptions that the residuals are normally distributed and have constant variance, leading to inaccurate results (McDonald, 1999).

To meet these challenges of modeling educational and psychological assessments, the common factor model has been extended in two main directions—multidimensional models that are capable of representing more complex relationships among multiple latent variables (e.g., higher-order and bi-factor models) and non-linear models that can be used with categorical observed variables (i.e., Item Response Theory models) (McDonald, 1982).

Higher-Order Models

The higher-order model (Jöreskog, 1971b) is an extension of the common factor model that is commonly used to represent the multidimensional construct structures of educational and psychological assessments. In this model, a second-order factor representing the overarching construct of interest is added above the typical first-order factors that represent subsets of items on the test designed to measure sub-parts of the overarching second-order factor (i.e., subscales). To use this model, the first-order factors must be substantively correlated and the second-order factors should be hypothesized to account for the variation among the first-order factors, in addition to the assumptions for a traditional unidimensional model (Wang & Wang, 2012).

Under the higher-order model there is only an indirect relationship between the items and the second-order construct—the items are indicators of their respective first-order factors, which are in turn indicators of the second-order factor (Cucina & Byle, 2017). Rijmen (2010) demonstrated that the higher-order model is formally equivalent to

the testlet model (Bradlow, Wainer & Wang, 1999; 2007), which is also commonly used to model interdependencies among groups of items sharing a common stimulus on an assessment (e.g., items set in a common context like the PSIs).

Higher-order models are useful for investigating the existence of a second-order factor and are currently the most commonly cited approach for modeling multidimensional assessments (Cucina & Byle, 2017; Reise, 2012; Toland et al., 2017). However, the higher-order model has several notable limitations. Primarily, it is not possible to separate the items' specific relationships with the first- and second-order factors because the relationship between the items and second-order factors are mediated by the first-order factors. Thus, the higher-order model is not useful for analyses aimed at determining the relative strength of the items' relationships with the general construct versus the sub-constructs (Reise, 2012; Toland et al., 2017). Also, recent comparisons of the higher-order model and a less commonly used alternative, the bi-factor model, suggest that the bi-factor model typically provides superior model fit (Cucina & Byle, 2017; Toland et al., 2017).

Bi-factor Models

The bi-factor model (Holzinger & Swineford, 1937; Holzinger & Harman, 1938) is another extension of the common factor model in which each item serves as an indicator of both the general construct or dimension that the instrument is designed to measure and one other specific dimension. However, unlike the higher-order model, each item has a direct relationship with both the general dimension and the specific dimension to which it was assigned, making it possible to identify the unique influence of the

general dimension and specific dimension on each item (Toland et al., 2017). The specific factors represent the variance common to the groups of items beyond the general factor (DeMars, 2013). The bi-factor model has been used to account for intended common content among groups of items on an instrument (e.g., mathematics content domains) and additional “nuisance” dependencies among items (e.g., item blocks) to obtain more meaningful factor scores for the construct of interest (DeMars, 2013; Toland et al., 2017).

Historically, the bi-factor model has been outshone by the higher-order model, but recent literature suggests that the bi-factor model is superior for evaluating the internal validity of tests or scales with groups of items (Cucina & Byle, 2017; DeMars, 2013; Reise, 2012; Toland et al., 2017). According to Toland et al. (2017), in addition to making it easier to interpret the direct influence of the general factor on each item and understand the relative importance of each factor, the bi-factor model has the necessary psychometric properties for determining interpretable scores on both the general and specific factors (DeMars, 2013) and allows for more seamless investigation of the influences of the general and specific traits on other variables (e.g., in a subsequent structural equation model) (Chen, West & Sousa, 2006). It has also been found to provide more accurate estimates of item parameters, person traits, and reliability than unidimensional models and other competing models for groups of items within an instrument including the higher-order model and testlet models (DeMars, 2006; Toland et al., 2017).

To date, the bi-factor model has been successfully used several times with TIMSS and PIRLS (IEA's Progress in International Reading Literacy Study) data despite characteristics of these data that commonly present challenges in latent variable modeling (i.e., the number of dimensions, number of items, grouping of items in booklets, and sampling design). For example, Rijmen, Jeon, von Davier, and Rabe-Hesketh (2014) used TIMSS 2007 data to compare the fit of two bi-factor models with the content and cognitive domains as the specific factors, as well as a tri-factor model, which simultaneously classified the achievement items to their content domain and more specific framework topic areas. Using the PIRLS 2006 data, Rijmen (2011) applied the bi-factor model to investigate the extent to which clustering of items around common reading passages and the items classifications by comprehension processes impacts measurement of reading ability.

Item Response Theory Models

Item Response Theory (IRT) models were developed to overcome the second issue faced in the modeling the underlying structure of educational and psychological assessments—that the observed variables are typically categorical, and therefore violate the assumptions of linear factor models. The IRT model addresses this by using a link function that transforms the probability of an observed response on a categorical variable into a more continuous variable that will not violate the model assumptions (McDonald, 1982). Rather than assuming a linear relationship between the items and factor scores, the log-odds (natural log of the odds ratio) of the probability of responding correctly to an item is used to link the observed responses to the latent variables, resulting in symmetric,

unbounded, outcome variables in the logit metric that can be linearly related to the latent traits. Using this approach, the residual variance is not estimated, but instead assumed to follow a logistic distribution with a known residual variance (de Ayala, 2009). The estimates produced are considered “test free,” meaning that they can be placed on the same latent continuum regardless of the specific subset of items the respondent answered, and “sample free,” meaning that the item parameters are not dependent on the group of respondents (de Ayala, 2009).

The IRT family of models can take into account the item difficulty (one-parameter, or 1PL model), item difficulty and discrimination (two-parameter, or 2PL model), or item difficulty, discrimination, and a guessing parameter approximating the probability of randomly selecting a correct response (three-parameter, or 3PL model) in estimating a respondents’ ability on the latent trait (de Ayala, 2009). These models also can be used with nominal items, partial credit items, and rating scales, as well as multidimensional extensions of the CFA model such as higher-order and bi-factor models. de la Torre & Song (2009) established the use of the higher-order IRT model approach. Gibbons and Hedeker (1992) established the use of the full information bi-factor model for binary data, or bi-factor IRT model, enabling its use with binary item response data. Rijmen (2011) established the use of this model with ordinal data.

Chapter 3: Methods

Developing the eTIMSS 2019 mathematics PSIs was a highly collaborative process primarily involving staff at the TIMSS & PIRLS International Study Center and a dedicated cadre of expert consultants. Software developers and programmers at the TIMSS & PIRLS International Study Center and IEA Hamburg developed the eTIMSS software and custom programmed the PSIs. The TIMSS 2019 National Research Coordinators (NRCs) from each participating country reviewed all of the test content and implemented the eTIMSS 2019 Field Test in their respective countries.

As the TIMSS 2019 mathematics coordinator, the author of this dissertation was responsible for guiding the development of the mathematics PSIs and was therefore highly involved in the end-to-end development process. This included drafting and refining the items and scoring guides consistent with suggestions from the expert consultants and measurement principles, facilitating review meetings, working with graphic designers and programmers, and participating fully in the extensive quality assurance work that was needed to ensure the PSIs came to fruition and were presented to students as intended.

Chapter 3 has three main sections. The first two sections—*PSI Development Methods* and *eTIMSS 2019 Field Test*—describe the cooperative efforts of those involved in developing the eTIMSS 2019 mathematics PSIs and conducting the eTIMSS 2019 Field Test. Particular attention is given to the methods and procedures designed to ensure test content validity and student response process validity. The third section, *Analysis Methods*, describes the additional methods the author used to evaluate the validity of the

mathematics PSIs. The work described in the third section, as well as Chapters 4 and 5 of this dissertation, are solely the author's contribution to the study.

PSI Development Methods to Promote Validity

Overview

In many ways, PSI development work in mathematics followed the standard TIMSS procedures for ensuring content validity. It began with defining the target construct and assessment specifications by establishing the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017). Then, mathematics and measurement experts began creating context-based sets of items that measured the mathematics content and cognitive domains described in the framework. However, because the PSIs involved a more innovative approach to assessing mathematics ability compared to the traditional TIMSS items and doing so by capitalizing on the digital assessment environment, PSI development required additional efforts to ensure that these innovative tasks provided valid measurement of the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017) and served their intended purpose of increasing the fidelity with which TIMSS scores represent mathematics ability.

From a measurement perspective, developing suitable problem contexts and questions to guide students through the PSI item sets, or tasks, was considerably more challenging than developing traditional achievement items. Because of their complexity, each PSI warranted even more numerous reviews than is regularly required by TIMSS. Particularly in the early stages of the transition to eTIMSS, development work also necessitated close collaboration with programmers to design the user interface,

interactive features, and enhanced item types for the tasks. Once a prototype for each task was established, operationalizing the PSIs for delivery to students required considerable front-end and back-end programming work and extensive quality assurance to make sure the tasks functioned as intended. Additionally, because the PSIs are so unique and were developed in tandem with the eTIMSS assessment systems, cognitive laboratories and a series of small-scale pilot tests were needed to try out the tasks and systems before large-scale administration.

TIMSS set the ambitious goal of developing around a dozen PSIs for the eTIMSS 2019 Field Test, or three to five tasks in each subject and grade. To meet this goal, development work on the mathematics and science PSIs began in March 2015, more than two years before item writing for the rest of the eTIMSS 2019 assessments. The following sections describe the methods used in developing the eTIMSS mathematics PSIs, from establishing the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017) to analyzing the results of the eTIMSS 2019 Field Test. A complete timeline of the PSI development milestones from March 2015 through September 2018 is provided in Appendix A.

Developing the TIMSS 2019 Mathematics Framework

The first step in every TIMSS assessment cycle is to identify and prioritize the mathematics content and skills that the assessment will measure. Because TIMSS is a trend study, the assessment framework cannot drastically change from cycle to cycle, but is routinely updated to keep up with fresh ideas and current information about curricula, standards, and instruction in mathematics education around the world (Mullis & Martin,

2017). For TIMSS 2019, the author of this dissertation conducted a domain analysis prior to updating the mathematics frameworks for the fourth and eighth grades, which primarily focused on reviewing countries' descriptions of their mathematics curricula in the *TIMSS 2015 Encyclopedia* (Mullis et al., 2016) and analyzing teachers' responses to a topic-by-topic survey about the frequency with which the mathematics content in the *TIMSS 2015 Mathematics Framework* (Grønmo, Lindquist, Arora & Mullis, 2013) was taught at the target grades. The author identified commonalities across countries' curricula as well as any widespread discrepancies between countries' curricula and the *TIMSS 2015 Mathematics Framework* (Grønmo et al., 2013) to detect topic areas that may need to be updated.

Using the *TIMSS 2015 Mathematics Framework* (Grønmo et al., 2013) as the foundation, consultants and staff at the TIMSS & PIRLS International Study Center reviewed and revised each mathematics topic area within each grade level with the goal of clearly describing reasonable and observable forms of evidence that can support the target claims about mathematics achievement. Consistent with previous versions of the framework, each mathematics topic area was stated in terms of measurable knowledge and skills to ensure that the framework provided a clear definition of the construct and clarity for item writers.

The draft of the updated mathematics framework first was reviewed in September 2016 by the TIMSS 2019 Science and Mathematics Item Review Committee (SMIRC), a group of international content experts that helped guide the development of the TIMSS 2019 achievement items. The draft was then revised and subsequently reviewed by

country representatives from each participating country (NRCs), at the 1st TIMSS 2019 NRC meeting in February 2017. Following the NRC meeting and another round of revision, the NRCs were asked to review the updated draft again and provide additional feedback via a topic-by-topic online survey at both the fourth and eighth grades. The *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017) was finalized after the first round of regular eTIMSS/paperTIMSS item writing in May 2017 at the 2nd TIMSS 2019 NRC meeting. This allowed TIMSS to verify that it was possible to write TIMSS items to measure all topic areas as described in the framework.

The *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017) includes three content domains at the fourth grade—number, measurement and geometry, and data—and four at the eighth grade—number, algebra, geometry, and data and probability. Each TIMSS content domain consists of multiple topic areas, which are each comprised of several topics that describe the specific competencies the assessments measure. The three TIMSS cognitive domains—knowing, applying, and reasoning—are the same at both grades. Each cognitive domain also consists of multiple cognitive processes that provide detailed description of the specific practices the assessments are designed to elicit. To ensure that the assessments provide appropriate coverage of mathematics ability, the framework specifies the target percentage of testing time allocated to each topic area and cognitive domain.

Initial Task Development

In March 2015, staff at the TIMSS & PIRLS International Study Center began collaborating with members of the SMIRC to start developing the PSIs. Several members

of the mathematics SMIRC were asked to work closely with TIMSS staff to develop the PSIs, which included providing initial ideas for the tasks and participating in a series of meetings with TIMSS staff and other experts to develop and refine the problem contexts and items. In the early stages of development, specifications for what constitutes a successful PSI were refined and elaborated upon to establish clear development goals and preliminary decisions were made about the user interface and available tools. This work also involved creating scoring guides and scorer training materials, informing the machine scoring specifications, and providing ideas for event data capture.

The author of this dissertation, together with the graphic designers and programmers at the TIMSS & PIRLS International Study Center, was primarily responsible for facilitating the development of the mathematics PSIs together with the user interface. The programmers at IEA Hamburg were primarily responsible for the complicated and time consuming work of preparing the PSIs to be administered to students via the eTIMSS assessment systems.

Expert Review

The PSI development process involved numerous rounds of expert review. Leading up to the field test, mathematics consultants and staff at the TIMSS & PIRLS International Study Center met a total of five times at Boston College and conducted countless online reviews to refine the tasks. Given the variety of challenges faced in developing the PSIs, this iterative and extended review process was critical for developing a cohesive series of achievement items for each problem solving context.

The SMIRC conducted an in-depth review of the mathematics PSIs at the 1st TIMSS 2019 SMIRC meeting in April 2017. The SMIRC focused on the alignment between the tasks and the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017) and the extent to which the technology in the tasks supported the intended response processes. The SMIRC also provided feedback about the cross-cultural appropriateness of the tasks.

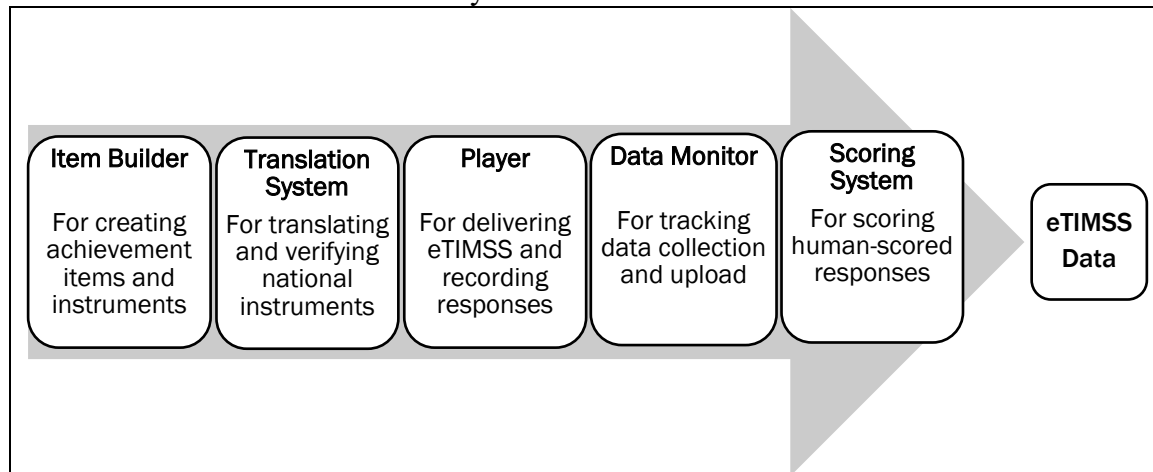
The NRCs reviewed the PSIs prior to the field test as well at their 3rd TIMSS 2019 NRC meeting in November 2017. In May 2018, NRCs were asked to provide additional feedback on the PSIs based on their experiences in the field test so that the TIMSS & PIRLS International Study Center could begin revising the PSIs for eTIMSS 2019 Data Collection as soon as possible. Staff at the TIMSS & PIRLS International Study Center reviewed all NRC comments, selected the PSIs for the eTIMSS 2019 assessments based on NRCs' recommendations, and began editing the selected tasks in June 2018. In July 2018, the SMIRC reviewed and further revised the mathematics PSIs at the 3rd TIMSS 2019 SMIRC meeting. One month later in August 2018 at their 5th meeting, the NRCs conducted a final review of all the eTIMSS 2019 achievement instruments, including the mathematics PSIs.

eAssessment Systems and Programming

Transitioning to eTIMSS also required developing a complex eAssessment infrastructure through which the eTIMSS assessments could be created, translated, delivered to students, and scored. IEA Hamburg began collaborating with staff at the TIMSS & PIRLS International Study Center on this extensive undertaking in January

2015 and development work continued through the start of main data collection in September 2018. Exhibit 3.1 presents the five components of the eTIMSS assessment system.

Exhibit 3.1: eTIMSS Assessment System



The eTIMSS Item Builder is a web-based application for creating the eTIMSS items and assembling assessment instruments. For eTIMSS 2019, it offered both traditional item formats (multiple-choice and constructed response) and several enhanced item formats—drop-down menus, selection, drag and drop, and sorting. The item builder included a variety of features for designing items, such as tools for uploading and adding text to images, creating tables, and previewing what the item will look like on a computer or tablet. It also contained the “assembler,” which was used to organize the items into test forms. With the exception of the PSIs, staff at the TIMSS & PIRLS International Study Center created all eTIMSS 2019 achievement items and instruments in the item builder (see next section, *Programming the PSIs*).

Once the eTIMSS 2019 international achievement instruments were complete, IEA Hamburg released the instruments to the online eTIMSS Translation System,

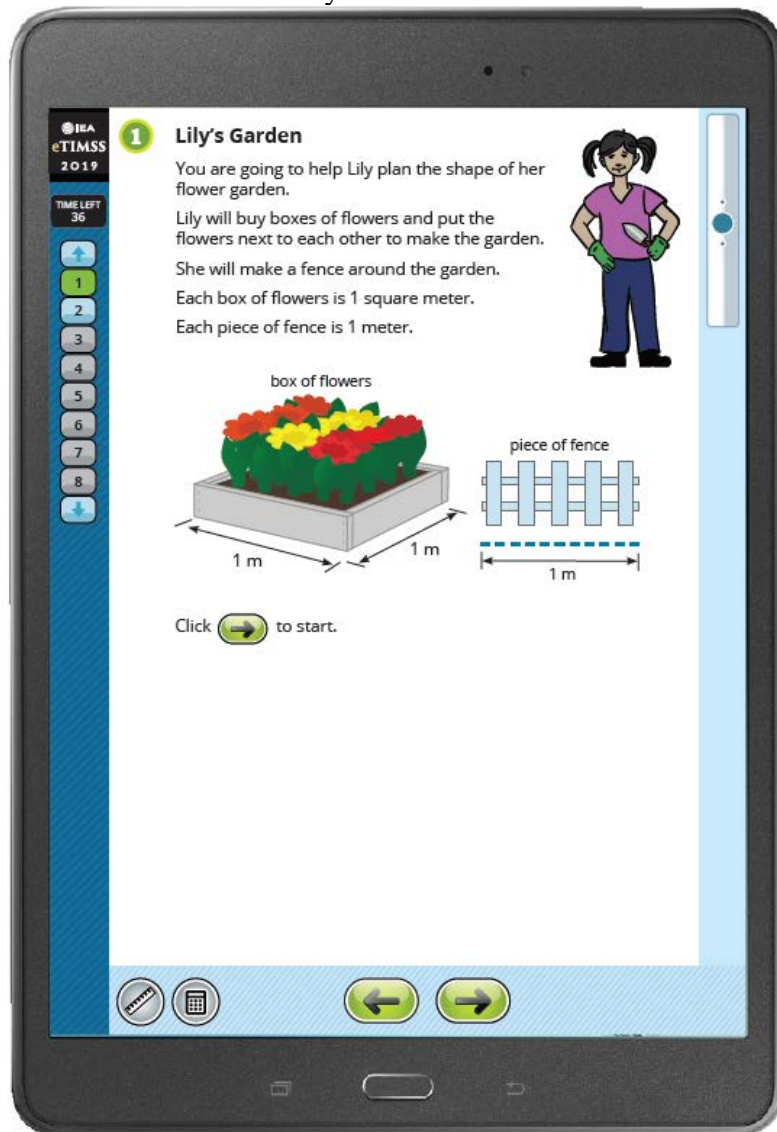
through which NRCs translated and adapted the items to their national language(s) of instruction. Translation verifiers at IEA Amsterdam and layout verifiers at the TIMSS & PIRLS International Study Center used the system to review countries' national instruments and provided comments until each countries' national instruments were finalized.

The eTIMSS Player is the software application used to deliver the assessment on computers and tablets. For eTIMSS 2019, it was compatible with a variety of devices and did not require an internet connection for test delivery. While the eTIMSS Player was running, it restricted access to all other programs to prevent distractions during testing. Exhibit 3.2 presents the user interface for the eTIMSS Player on a tablet. On a PC, the same dimensions of the rectangular screen were preserved and a blue background was added to the left and right of the player interface to fill the rest of the computer screen.

The item number was displayed in the top left corner of the screen, as well as a timer that showed the number of minutes the student had left to complete a part of the assessment. On the left side of the screen, there were numbered buttons for each item in the part of the assessment the student was working on. These buttons were initially grey, then turned green to indicate the current screen the student was on and dark blue once a screen had been visited. On the bottom of the screen there were green 'back' and 'next' arrows that students could use to move through the assessment item by item. On the right side of the screen there was a scroll bar that students could use to see any parts of items that extended below the bottom of the window. There were also two buttons for the eTIMSS tools—the eTIMSS on-screen ruler and the calculator—at the bottom of the

screen that were made available to students on a screen-by-screen basis. At both grades, the ruler tool was available for a small number of mathematics items involving measuring. Consistent with TIMSS policy, the calculator tool was not available at the fourth grade, but was available for all items at the eighth grade. When available, the icons for the tools appeared in blue; when the tools were activated, the icons turned orange.

Exhibit 3.2: eTIMSS Player Interface



Introduction screen from an example fourth grade mathematics PSI task, *Lily's Garden*, in the eTIMSS Player.

Students logged in to the eTIMSS Player with a unique username and password provided by the test administrator. The software saved students' responses directly on the testing device or on a central server computer, depending on the method of administration. At the end of a testing session, test administrators uploaded the data to the IEA's server and used the online eTIMSS Data Monitor to check that the data were captured and uploaded correctly.

All responses to constructed response items that required human scoring were sent to the IEA's Online Scoring System that was used to distribute student responses to scorers and enter a score code for each response. The system enabled national scoring coordinators to systematically assign responses to members of the scoring team (e.g., by item, block, or language of test), monitor scorers' progress, and review responses that scorers flagged with questions. Scorers viewed student responses in the system and selected from the available score codes to classify the response according to the scoring guide. The scoring system also facilitated all activities TIMSS requires to assess scoring reliability (within-country, cross-country, and trend) and could be used to train scorers by adding example student responses in the system for practice.

Programming the PSIs

To allow for a broader variety of interactive features beyond the standard eTIMSS item types offered in the item builder, each PSI was individually programmed by staff at the TIMSS & PIRLS International Study Center and IEA Hamburg. Programming the PSIs was an extremely time consuming and resource intensive process because of the wide variety of unique features required for each task. Also, PSI development work

coincided with the development of the eTIMSS infrastructure, making it challenging to ensure the tasks were compatible with the rest of the assessments, as the systems were constantly evolving.

Each mathematics PSI began as a screen-by-screen outline of the task including the text, example images, and notes for how the proposed interactive features should function. Once an outline was thoroughly reviewed and deemed ready for programming, the author of this dissertation collaborated with graphic designers at the TIMSS & PIRLS International Study Center to create “storyboards,” laying out each screen in the eTIMSS Player interface on paper, and prepared detailed “coding notes” to explain the desired functionality to the programmers. The programmers then began creating prototype versions of the tasks, which typically resulted in further revisions to both the content and functionality of the tasks as the programmers determined what was feasible. Once a PSI was mostly operational, it was made available for quality assurance in IEA Hamburg’s Quality Assurance System, then eventually merged into the eTIMSS Item Builder to be assembled into test instruments along with the regular eTIMSS items.

Cognitive Laboratories, Pilot Testing, and Observations

During the development process, cognitive laboratories and a series of pilot tests in the eTIMSS countries were conducted to gain insight into students’ interactions with the PSIs and test the functionality of the eTIMSS assessment systems. This strand of development work provided critical insight into the usability of the innovative item types and eTIMSS interface, the amount of time for students to complete each task, and the approximate difficulty of the tasks. It included cognitive laboratories in August 2015, the

eTIMSS prePilot in three countries in October 2016, and the eTIMSS Pilot / Item Equivalence Study involving 25 countries in May 2017. In March through May 2018, TIMSS conducted a full-scale field test with 31 countries at the fourth grade and 22 countries at the eighth grade participating in eTIMSS.

Following each study, improvements were made to both the PSIs and eTIMSS assessment systems to ensure that the novel aspects of the eTIMSS experience were eliciting the intended responses from students and enhancing measurement of mathematics knowledge and skills as intended.

Cognitive Laboratories

Staff at the TIMSS & PIRLS International Study Center partnered with the American Institute for Research (AIR) to conduct cognitive laboratories in the very early stages of the transition to eTIMSS (August 2015). The goal of this study was to investigate two aspects of digital assessment that would inform next steps in eTIMSS development: 1) students' interactions with drafts of the first PSIs, and 2) students' experiences with the eTIMSS interface.

The TIMSS & PIRLS International Study Center prepared two prototype PSIs—*Lily's Garden* for fourth grade mathematics and *Pepper Plants* for eighth grade science—and a set of TIMSS trend items at each grade on tablets, which was the anticipated mode for eTIMSS at this point. All items were designed for students to respond using a stylus (or finger), including those that required a written response. In the first version of the eTIMSS Player students were able to write or draw anywhere on the tablet screen, mimicking the paperTIMSS experience to the extent possible.

Staff at the TIMSS & PIRLS International Study Center provided a list of research questions, from which AIR developed interview protocols that incorporated a think aloud aspect and reflective aspect. During the interviews, students explained their thoughts while engaging with the items, providing insight into how the PSI format and eTIMSS interface could be improved.

AIR conducted the interviews with a purposive sample of 32 fourth and eighth grade students from the greater Washington, D.C. area. Interested participants were screened to ensure a range of mathematics and science ability, frequency of computer use for educational purposes, socioeconomic background, and a balance of females and males. At the fourth grade, seven students completed *Lily's Garden* and eight students completed the trend items. At the eighth grade, eight students completed *Pepper Plants* and nine students completed the trend items.

Following the interviews, AIR prepared a report to address each of the TIMSS & PIRLS International Center's research questions. In November 2015, consultants and staff at the TIMSS & PIRLS International Study Center met at Boston College to review the results, revise the prototype PSIs, and continue developing additional tasks based on new insights from this study. The results from the cognitive laboratories prompted several substantial revisions to the PSIs and the eTIMSS interface.

Reconsidered the stylus/finger approach. The reports indicated that students at both grades experienced difficulties trying to use styluses or fingers to write or draw their responses. The technology was difficult to control and not precise enough. Students relied heavily on scratch paper and reported that they wrote less on the tablet than they would

have on paper. Based on these results, the eTIMSS Player was updated to provide students with the option of using an on-screen keyboard to respond to constructed response items requiring a written response. Efforts were also made to improve the technology for free-hand drawing and writing with a stylus/finger in hopes of making it a viable response mode.

Eliminated tutorial videos. The *Lily's Garden* prototype began with a video tutorial to explain the problem situation and teach students how to use the interactive response spaces in the task. The reports indicated that many students appeared to be confused during the tutorial or unsure of how to proceed when the tutorial ended. However, most students were able to successfully interact with the enhanced response spaces despite the initial confusion, indicating that the elaborate videos were unnecessary. Following the cognitive laboratories, the *Lily's Garden* tutorial video was replaced with a static screen to provide a more streamlined explanation of the task and its features and future PSIs followed the same approach.

Added a 'back' button. In both prototype PSIs it was only possible to move forward through the tasks so that students could not go back and change their answers when the answer to an item was given away on a later screen. The majority of students reported that they wanted to return to a previous screen to check their work or re-read information to help them understand the problem. These comments, coupled with the belief that students taking the PSIs should have the same ability to freely navigate as students taking the regular eTIMSS items, resulted in the addition of a 'back' button within all PSIs. This change required that each PSI be carefully designed to avoid giving

away answers to questions on later screens, but eliminated students' cause for hesitation before moving on to the next screen and made it possible for students to review their final answers, like they are traditionally encouraged to do when taking TIMSS.

Simplified functionality of novel response spaces. Several of the enhanced item formats in the prototype PSIs provided students with more than one method of conveying their response. For example, in the *Lily's Garden* prototype there were two available methods for creating a fence outline of a garden on a grid—dragging pieces of fence onto the grid and tapping on the grid lines to make pieces of fence appear. This flexibility was intended to make the novel response spaces intuitive to use for as many students as possible, but the reports indicated that offering multiple ways to respond to an item was more confusing and cumbersome than helpful. Following the cognitive laboratories, additional efforts were made to keep interactive elements in the PSIs as simple and user friendly as possible.

eTIMSS prePilot

In September 2016, four fourth grade mathematics PSIs and three eighth grade mathematics PSIs were piloted in a standard eTIMSS testing situation for the first time. Considerable advances had been made in both task and eTIMSS Player development since the cognitive laboratories, enabling TIMSS to try out a broader variety of interactive features and enhanced item types in this administration. Piloting the PSIs with a larger group of students also helped consultants and staff at the TIMSS & PIRLS International Study Center gauge the item and overall task difficulty, as well as the average time to complete each task.

Based on countries’ feedback on the initial plans for eTIMSS, it was decided to offer the assessment on both tablets and PCs to accommodate a wider range of devices and support more countries in participating. For the prePilot, the standard device keyboard was made available for all constructed response items requiring a text-based response and a drawing tool students could use with a stylus, finger, or mouse was only enabled for items involving drawing or showing work. For constructed response items requiring a numeric answer, students were provided with an on-screen number pad to enter their responses that included the digits 0 to 9, a decimal point, and enter and backspace buttons.

The eTIMSS 2019 prePilot instruments were comprised of a total of six item blocks each containing 12 to 15 items at the fourth grade and 14 to 16 items at the eighth grade. There were three mathematics blocks and three science blocks at each grade—two blocks comprised of PSIs and one block of regular TIMSS trend items converted to digital format. Exhibit 3.3 presents the contents of the six blocks at each grade.

Exhibit 3.3: eTIMSS 2019 prePilot Blocks

Block	Grade 4	Grade 8
M01	TIMSS 2015 Mathematics Block	TIMSS 2015 Mathematics Block
M02	<i>Lily’s Garden and Robots</i>	<i>Building and Robots</i>
M03	<i>Little Penguins and Blue and White Picture</i>	<i>Clothing Store</i>
S01	<i>Farm CSI</i>	<i>Pepper Plants</i>
S02	TIMSS 2015 Science Block	TIMSS 2015 Science Block
S03	<i>Sugar and Water and Magnet Train</i>	<i>Sunken Ship</i>
Blocks M01 and S01 were comprised of mathematics trend items and science trend items, respectively. Blocks M02 and M03 were comprised of mathematics PSIs and blocks S02 and S03 were comprised of science PSIs.		

At both grades, the six blocks were used to create three test forms, referred to as “block combinations” for eTIMSS, each comprised of two mathematics blocks and two

science blocks. Exhibit 3.4 shows the three block combinations that were used at both grades. Each block appeared in two block combinations.

Exhibit 3.4: eTIMSS 2019 prePilot Block Combinations

Block Combination	Part 1		Part 2	
BC01	M01	M02	S01	S02
BC02	S02	S03	M02	M03
BC03	M03	M01	S03	S01

The prePilot was conducted in three English-speaking countries with experience in conducting digital assessments: Australia, Canada, and Singapore. Each country selected 2 to 4 classes at each grade to participate and made efforts to include students with a range of mathematics and science ability. This sample yielded approximately 100 responses per item at both the fourth and the eighth grade.

Students' responses to the PSIs, the TIMSS & PIRLS International Study Center and IEA Hamburg's experiences in delivering eTIMSS in a standard testing situation, and detailed reports from the Australian Council for Educational Research (ACER) and Singapore NRCs provided more ideas for improvement. The results of the prePilot were reviewed both from a content perspective, by mathematics consultants and staff at the TIMSS & PIRLS International Study Center, and an operational perspective, by staff at the TIMSS & PIRLS International Study Center and IEA Hamburg. Following the prePilot, more changes were made to both the PSIs and the eTIMSS assessment systems to prepare for the field test.

Reduced the item difficulty. A substantial number of items in the PSIs had very low percentages of correct responses and high omit rates, suggesting that the prePilot versions of the tasks were too difficult for the target grade levels. Following the prePilot,

the mathematics consultants revised the most difficult items within each task to be more grade appropriate by simplifying the numbers or adding more scaffolding and confirmed that all items strictly adhered to the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017).

Reduced the reading load. The prePilot PSIs included significantly more text than regular eTIMSS items and test administrators observed students becoming restless or less engaged as they worked through the PSIs because of the heavy reading load. The prePilot data further reinforced this observation with the higher omit rates on screens with more text and screens towards the end of the PSI blocks, by which students were likely fatigued. Following the prePilot, the mathematics consultants and staff at the TIMSS & PIRLS International Study Center substantially reduced the reading load in all PSIs.

Updated the number pad design. There were no issues capturing students' responses via the number pad, but reports indicated that some students appeared to be frustrated with this feature and suggested that this was due to the unfamiliar arrangement of the buttons. Following the prePilot, the number pad was updated to match the layout of a standard keyboard and a negative button was added, as well as the capability to respond with a fraction.

Continued improving the eTIMSS Player and Data Monitor. Overall, the testing sessions went smoothly, with the exception of several students being suddenly logged out during the test and a small number of computer crashes that resulted in a loss of data. Also, due to issues with uploading the data saved on USB sticks, the test

administrators in Singapore were only able to save the data for approximately half the students who completed the prePilot on a PC. Following the prePilot, staff at IEA Hamburg continued refining and testing the eTIMSS assessment systems to minimize such issues in future administrations.

eTIMSS Pilot / Item Equivalence Study

The eTIMSS Pilot / Item Equivalence Study was conducted in March through May 2017 for the purposes of examining the equivalence of the TIMSS trend items in digital and paper format and giving countries an opportunity to practice using the eTIMSS assessment systems on a relatively large scale (Fishbein, Martin, Mullis & Foy, 2018). Twenty-five countries participated—24 at the fourth grade and 13 at the eighth grade—with a sample of 800 students at each grade.

Participating in the eTIMSS Item Equivalence Study involved translating the trend items via the eTIMSS Translation System, preparing devices to be compatible with the eTIMSS Player, scoring constructed response items via the IEA’s Online Scoring System, and checking that data were uploaded to the IEA’s servers using the eTIMSS Data Monitor. The achievement instruments used for the study only included trend items (no PSIs), but prompted critical updates to the eTIMSS assessment systems that improved the entire assessment.

Expanded the eTIMSS manuals. Based on country feedback, staff at the TIMSS & PIRLS International Study Center expanded the survey operations and procedures manuals and test administration script to better support NRCs, school coordinators, and test administrators in conducting eTIMSS. This included adding specific instructions for

a wider variety of digital devices and more detail about managing the device settings, such as the default keyboard and the “autocorrect” feature on tablets. Directions for trouble shooting common issues that arose in the Pilot and tips from NRCs for ensuring smooth administration, such as a reminder to charge laptops between sessions, also were added.

Improved the eTIMSS Translation System. The version of the eTIMSS Translation System used in the eTIMSS Item Equivalence Study offered less flexibility in translating and adapting achievement items than the paper-based methods TIMSS participants had become accustomed to, causing frustration among participating countries and requiring extensive support from IEA Hamburg in preparing national instruments. This version of the system did not allow for the number pad to be translated, presenting issues for countries using a decimal comma, or for any general translations to be applied (e.g., item numbers, letters in multiple-choice answer options), resulting in tedious work for several countries. Also, some adaptations of mathematical symbols were unrecognizable by the system and had to be adjusted on a country-by-country basis. Further, countries experienced difficulties positioning translated text around images, particularly when translating the assessment into right-to-left formatted languages. Following the eTIMSS Item Equivalence Study, IEA Hamburg made significant improvements to the eTIMSS Translation System to address all of these issues and further developed the user interface to facilitate a smoother translation process.

Continued improving and testing the eTIMSS Player. The eTIMSS Player used in the eTIMSS Item Equivalence Study was the most advanced yet, but there were

still issues that needed to be addressed before full-scale data collection. In response to countries' reports of sporadic system crashes, freezes, or items not functioning correctly on particular devices, IEA Hamburg began even more extensive testing of all software and item types on a variety of devices and in multiple languages. IEA Hamburg also worked on improving the “lock” feature to prevent students from opening other programs and applications during testing after several countries reported this was an issue. Additionally, a new feature was added to the eTIMSS Data Monitor to indicate directly on the testing device when data are successfully uploaded to the IEA server to better support test administrators in managing their responsibilities and minimize missing data.

Eliminated the free-hand drawing tool. During the eTIMSS Item Equivalence Study students were still able to draw freely with their mouse, finger, or stylus to write their answers or show their work. Consistent with the cognitive laboratories and prePilot, countries' reports indicated that students experienced difficulties using this approach, particularly on a PC. Because of the variety of devices being used in testing, capturing students' responses for scoring also proved to be challenging. In some cases, it was not possible to accurately re-create students' responses in the IEA's Online Scoring System and responses to these items could not be scored. Although drawing and showing work are an important part of mathematics assessment, the approach of free-hand drawing was eventually determined to be infeasible for eTIMSS in 2019 and alternative enhanced item formats were explored.

Informed the development of eTIMSS items. The results of the eTIMSS Item Equivalence Study indicated that recreating the trend items in digital format introduced a

mode effect, with the mathematics items being overall more difficult in eTIMSS than on paper (Fishbein et al., 2018). Item-by-item analysis of the results suggested that in addition to items involving free-hand drawing or showing work, items on screens with excessive amounts of scrolling and text-based constructed response items with insufficient space to type exhibited the greatest mode effects. Staff at the TIMSS & PIRLS International Study Center kept these issues in mind in proceeding with eTIMSS 2019 item development and worked to minimize features that were found to be associated with mode effects.

Developing Scoring Guides for Constructed Response Items

Following standard TIMSS procedures, scoring guides and distracter rationales for each item were developed concurrently with the PSIs and included in all expert reviews. In January through February 2018, consultants and staff at the TIMSS & PIRLS International Study Center convened at Boston College to review the PSI scoring guides in light of new information about data capture and machine scoring capabilities and prepare scorer training materials for several complex, human-scored constructed response items.

The PSI scoring guides used the same TIMSS generalized scoring guidelines and two-digit diagnostic scoring system that have proven successful in ensuring a high degree of scorer agreement in previous assessment cycles. Under this system, the first digit indicates the degree of correctness of the response (score points) and the second digit provides diagnostic information (e.g., a specific method for solving a problem or a common misconception). The eTIMSS assessments at both grades include both

dichotomous items worth one score point (scored as 1=correct, 0=incorrect) and polytomous items worth two score points (scored as 2=fully correct, 1=partially correct, 0=incorrect). Transitioning to digital assessment made it possible to machine score most constructed response items types, including all items using the number pad, eTIMSS components, and most of the customized item types designed specifically for the PSIs. For the mathematics assessment, this comprised the majority of the constructed response items in the eTIMSS 2019 Field Test.

The same basic approach was used in developing scoring guides for machine- and human-scored items, but more attention was given to different aspects of the guides depending on how the item was to be scored. For machine-scored items, the scoring guides served as the basis for machine scoring specifications, so it was important that the guides clearly defined each code in such a way that it could be accurately applied without human judgement of student responses. For example, many mathematics scoring guides for items using the number pad included a range of acceptable values for each score code to account for rounding in computations or specifications for the acceptable number formats of a response (e.g., whether fractions and decimals may both be accepted). Most human-scored mathematics items involved either an explanation or justification of an answer or, at the eighth grade, an algebraic equation or expression typed on a keyboard. The description of each score code for a human-scored item typically included a general statement describing the required qualities of a response in the category, followed by several examples of student responses that would receive the code.

eTIMSS 2019 Field Test

Overview

In preparation for data collection, TIMSS routinely conducts a full-scale field test for the purposes of evaluating the measurement properties of the item pool and practicing operations procedures to ensure smooth administration for the main study (Mullis et al., 2016). TIMSS field tests approximately one and a half times the number of items needed for the final instruments to allow for the best items to be selected for data collection. Main data collection for eTIMSS 2019 was still underway at the time of this dissertation, so the data collected in the eTIMSS 2019 Field Test was used to conduct a preliminary analysis of the measurement properties and internal structure of the eTIMSS mathematics assessments.

The eTIMSS mathematics field test instruments were comprised of 174 items (127 regular and 47 PSI) at the fourth grade and 201 items (158 regular and 43 PSI) at the eighth grade. With the exception of the PSIs, each of these items was also field tested in paper format (paperTIMSS).

Achievement Instrument Design

The regular field test item pool for each subject and grade was divided into 10 unique, balanced, blocks of items each consisting of 12 to 15 items at the fourth grade and 14 to 16 items at the eighth grade. The regular blocks at each grade were organized into five block combinations for eTIMSS and five booklets for paperTIMSS, each of which was comprised of two mathematics blocks and two science blocks. The regular blocks were distributed across these block combinations/booklets using an incomplete

and un-rotated design in which each block appeared in a single block combination/booklet. These five block combinations/booklets were designed to be identical in content across eTIMSS and paperTIMSS, with the only differences being in the response mode (e.g., a drag and drop item in eTIMSS may become a matching item in paperTIMSS).

For eTIMSS, the field test instruments also included three additional block combinations of PSIs for each subject and grade. These block combinations employed a balanced incomplete block design in which the three PSI blocks for each subject and grade appeared twice—once with each of the other PSI blocks—and were rotated across the combinations to account for potential position effects. Exhibit 3.5 shows the regular block combination/booklet design for the five block combinations/booklets used in both eTIMSS and paperTIMSS (block combinations/booklets 1–5) as well as the PSI block combinations used exclusively for eTIMSS (block combinations 6–8).

Exhibit 3.5: eTIMSS 2019 Field Test Block Combinations/Booklets

Block Combination/ Booklet	Part 1		Part 2	
1	ME01	ME02	SE01	SE02
2	SE03	SE04	ME03	ME04
3	ME05	ME06	SE05	SE06
4	SE07	SE08	ME07	ME08
5	ME09	ME10	SE09	SE10
6	MI01	MI02	SI01	SI02
7	SI02	SI03	MI02	MI03
8	MI03	MI01	SI03	SI01

Blocks beginning with “ME” and “SE” are regular eTIMSS mathematics and science blocks, respectively. Blocks beginning with “MI” and “SI” are mathematics PSI blocks and science PSI blocks, respectively.

Each student participating in the field test completed one paperTIMSS booklet or eTIMSS block combination. For all TIMSS block combinations/booklets the total testing

time was 72 minutes at the fourth grade and 90 minutes at the eighth grade. At both grades, students spend half this time completing the first two blocks (Part 1), had a short break, then completed the second two blocks (Part 2). To accommodate this design, several of the PSI blocks consisted of a single task, while others were comprised of two tasks, depending on the number of questions per task. In all, the eTIMSS 2019 Field Test included five fourth grade mathematics PSIs and four eighth grade mathematics PSIs.

The TIMSS & PIRLS International Study Center provided international English versions of the field test achievement instruments, which countries translated and adapted via the eTIMSS Translation System to create their own national instruments. All national instruments underwent translation and layout verification to ensure international comparability. Prior to each countries' testing window, IEA Hamburg provided the country with a draft eTIMSS Player containing their national achievement instruments for testing and addressed any issues on a case-by-case basis until each player was approved for administration.

eTIMSS Student Questionnaire

After completing the achievement items, students taking eTIMSS were asked to respond to the eTIMSS Student Questionnaire. The questionnaire asked students to report the extent to which they liked taking the test on a computer, experienced difficulties responding to items or with their device, the frequency with which they use computers at school, and beliefs about their computer skills. Exhibit 3.6 shows the questions from the eTIMSS Student Questionnaire measuring students' liking and difficulties taking the test on a computer/tablet that were considered in this dissertation.

Exhibit 3.6: eTIMSS Student Questionnaire Items Measuring Student Enjoyment and Difficulties Taking the Test on a Computer or Tablet

1.

A. Did you like that this test was on a computer or tablet?

1 = I liked it a lot

2 = I liked it a little

3 = I didn't like it very much

4 = I didn't like it at all

B. Did you have any of these difficulties?

1 = Yes

2 = No

a) It was hard to type

b) I had trouble using the number pad

c) Objects were hard to drag

d) There was no good place to work out my answers

e) The computer or tablet was slow

f) I had to start my test over because of a computer or tablet problem

Source: eTIMSS Student Questionnaire in the eTIMSS 2019 Field Test, fourth and eighth grades.

Sample

In total, 63 countries and 10 benchmarking entities participated in the eTIMSS/paperTIMSS 2019 Field Test, and slightly more than half of these countries administered the digital assessment. At the fourth grade, 31 countries and 6 benchmarking entities participated eTIMSS, and at the eighth grade, 22 countries and 5 benchmarking entities participated. Exhibit 3.7 shows the list of countries that participated in the eTIMSS 2019 Field Test.

Exhibit 3.7: Countries in the eTIMSS 2019 Field Test

Austria (4)	Israel (8)	Spain (4)
Canada (4)	Italy (4 and 8)	Sweden (4 and 8)
Chile (4 and 8)	Japan (4 and 8)	Turkey (4 and 8)
Chinese Taipei (4 and 8)	Korea (4 and 8)	United Arab Emirates (4 and 8)
Croatia (4)	Lithuania (4 and 8)	United States (4 and 8)
Czech Republic (4)	Malaysia (8)	
Denmark (4)	Malta (4)	Benchmarking Participants
England (4 and 8)	Netherlands (4)	Ontario, Canada (4 and 8)
Finland (4 and 8)	Norway (4 and 8)	Quebec, Canada (4 and 8)
France (4 and 8)	Portugal (4)	Moscow, R. Fed. (4 and 8)
Georgia (4 and 8)	Qatar (4 and 8)	Madrid, Spain (4)
Germany (4)	Russian Federation (4 and 8)	Abu Dhabi, UAE (4 and 8)
Hong Kong (4 and 8)	Singapore (4 and 8)	Dubai, UAE (4 and 8)
Hungary (4 and 8)	Slovak Republic (4)	

Grade(s) of participation appear in parentheses. Countries that administered the paper version of the field test are not listed. Benchmarking participants were not included in analysis.

For both the field test and main data collection, TIMSS uses a two-stage random sampling design to ensure that data collected from a sample of students provides accurate representation of all students in the designated grade in each country. Leading up to the eTIMSS 2019 Field Test, sampling experts from Statistics Canada and IEA Hamburg worked with NRCs to define the target population in their country and specify the necessary information for establishing a sampling plan. Once a country's target population was defined, a sample of schools was randomly selected in the first stage, then one or more intact classes of students within each of the sampled schools were selected in the second stage (LaRoche, Joncas & Foy, 2016). TIMSS requires that countries diligently document coverage and participation rates to ensure that any selection bias introduced during sampling can be appropriately considered in analysis and reporting.

For the eTIMSS 2019 Field Test the sample size requirement was 200 students per block combination, or approximately 25 to 40 schools with two classes sampled per

grade (LaRoche, 2017). The block combinations/booklets were pre-assigned to students using the TIMSS within-school sampling software (WinW3S) to ensure that the sample of students that completed each instrument in each country is approximately equivalent in terms of student ability (Martin, Mullis & Foy, 2017).

Exhibit 3.8 provides a summary of the number of mathematics items and participants in the eTIMSS 2019 Field Test at the fourth and eighth grades.

Exhibit 3.8: Summary of eTIMSS 2019 Field Test Items and Participants

	Grade 4	Grade 8
Mathematics Items		
Regular Items	127	158
PSI Items	47	43
Total	174	201
Participants		
Countries	31	22
Benchmarking Entities	6	5
Schools	2,163	1,403
Students	43,293	31,116

Field Test Data Collection

Countries were offered three methods for delivering the eTIMSS 2019 Field Test—1) individual PCs with the eTIMSS Player on USB sticks, 2) individual tablets with the eTIMSS Player software installed, or 3) a server method, using a central PC or Chromebook as a local server that delivers the eTIMSS Player to students’ PCs/Chromebooks via the school’s Local Area Network (LAN). To minimize technical difficulties in running the eTIMSS Player and uploading the data, TIMSS provided minimum requirements for screen resolution, operating system, processor speed, memory, USB ports, and system font sizes for both computers and tablets. Countries were equipped with detailed manuals on preparing the devices for test administration,

running the eTIMSS Player, and uploading the eTIMSS data after a testing session.

Countries collected data between March and May 2018.

Timing Data

In addition to collecting student responses to items, the eTIMSS Player collected timing data indicating the number of seconds each student spent on the screens they encountered in the eTIMSS 2019 Field Test. In instances that a student visited a screen multiple times, the time on screen was calculated as the total number of seconds the student spent on the screen across all visits. IEA Hamburg provided Microsoft Excel files containing the average and median number of seconds students spent on each screen by country to the TIMSS & PIRLS International Study Center in June 2018.

The timing data presented new opportunities to investigate potential position effects associated with the different blocks of items. The position of the block in the block combination could impact the measurement properties of the items as well as student engagement and motivation. Also, by using time on task as a proxy for student effort, the author of this dissertation further evaluated the response process validity and internal structure of the tasks according to their cognitive domain classification and position in the block combinations (further explained in *Analysis Methods*).

Scoring Constructed Response Items

The data analysis team at the TIMSS & PIRLS International Study Center drafted the machine scoring specifications for each machine-scored constructed response item in the eTIMSS 2019 Field Test based on the scoring guides. These specifications linked the raw data that was produced from the eTIMSS Player to the definition of each code in the scoring guide for acceptable or unacceptable responses. Drafting the machine scoring

specifications involved testing each item in the eTIMSS Player, reviewing the output, then writing rules in terms of the output to classify all possible responses to a code in the item's scoring guide. The specifications included conventions for naming output variables from the eTIMSS Player, rules for processing numeric input and responses to enhanced item types, and rules for deriving scores for items with multiple parts.

The scoring unit at IEA Hamburg reviewed all specifications and provided feedback on an item-by-item basis, resulting in several rounds of revision until the rules for all items in the field test were clarified. The scoring unit at IEA Hamburg then applied the scoring rules for all machine-scored items and the data analysis team at the TIMSS & PIRLS International Study Center independently replicated the results to validate the scoring.

The TIMSS 2019 NRCs and their scoring supervisors received scoring training for the most complex human-scored constructed response items in the field test in March 2018, as part of the 4th TIMSS 2019 NRC meeting. This training included one item from a fourth grade mathematics PSI and six items from the eighth grade mathematics PSIs. The goal of this training was to ensure that the scoring guides for all human-scored items were applied consistently within and across countries. The training materials consisted of 8 to 12 student responses to illustrate the codes in the scoring guide (example responses), followed by 8 to 12 student responses without pre-assigned score codes to be used as practice during the training sessions (practice responses). At the training sessions, the trainers explained the purpose of each item and read it aloud. The trainer then described the scoring guide, explaining each category and the rationale for the score given to each

example paper. The country representatives were then given time to score the practice papers to practice making distinctions among categories. The correct codes for each practice paper were then reviewed, any inconsistencies in scoring were discussed, and, as necessary, the scoring guides were clarified and sometimes categories were revised.

Feedback from NRCs

In May 2018, the TIMSS & PIRLS International Study Center asked NRCs to provide feedback on the PSIs in the field test to facilitate an early start on selecting and making improvements to the PSIs that would move forward to main data collection. The NRCs provided a substantial amount of information about students' interactions with the PSIs and suggested specific revisions to improve the content and functionality of the tasks. The feedback also included new ideas for improving the eTIMSS directions, test administrator manuals, and assessment systems. The author of this dissertation prepared a summary of the NRCs' feedback that was used by staff at the TIMSS & PIRLS International Study Center to inform improvements for data collection.

The NRCs also reported on their experiences conducting the eTIMSS 2019 Field Test via the Field Test Survey Activities Questionnaire that was made available to countries in April 2018. The questionnaire included questions about NRCs' experiences preparing national instruments, conducting testing sessions, participating in quality control and monitoring activities, scoring constructed response items, and submitting data. The survey operations and procedures team at the TIMSS & PIRLS International Study Center prepared a question-by-question summary of results that was also used to guide improvements for data collection.

Item Review

Following field test administration and scoring, IEA Hamburg reconciled inconsistencies within and across countries' data and sent it to the TIMSS & PIRLS International Study Center for analysis. For reviewing field test data for items, TIMSS primarily uses classical item statistics, including item difficulty (average percent correct), item discrimination (point biserial-correlations) and missing rates (not applicable, omitted, and not reached) to evaluate the measurement properties of each item (Foy, Martin, Mullis, Yin, Centurino & Reynolds, 2016). These item statistics were calculated as follows:

- Item Difficulty – the average percent correct on an item. For 1-point items, it is the percentage of students providing a fully correct response; for 2-point items, it is the average percentage of points.
- Item Discrimination – the correlation between the response to an item and the total score on all items administered to a student (point-biserial correlation).
- Percent Omitted – the percentage of students who reached the item, but did not provide a response (not reached items are excluded from the denominator in calculating this percentage).
- Percent Not Reached – the percentage of students who were administered the item, but did not reach the item in the block combination/booklet. An item is designated “not reached” when the item itself and the item immediately preceding it were not answered and no subsequent items in the part of the block combination/booklet were attempted.

In June 2018, staff at the TIMSS & PIRLS International Study Center selected the mathematics items for data collection based on these item statistics and the test content. For the PSIs, the recommendations of the NRCs were confirmed by the field test results. The selected items were then reviewed by both the SMIRC and NRCs before the data collection instruments were finalized in August 2018.

Analysis Methods

Overview

Because eTIMSS 2019 Data Collection was still in progress at the time of this dissertation, each research question about the validity of the mathematics PSIs was answered using the most relevant sources of information currently available. Data collection began in the Southern Hemisphere in September 2018 and continued in the Northern Hemisphere through June 2019.

The content validity of the mathematics PSIs was evaluated based on the mathematics items in the final eTIMSS 2019 achievement instruments that were used in main data collection. These achievement instruments were designed to meet the content and cognitive specifications in the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017) and would provide the achievement data for the TIMSS 2019 International Reports in Mathematics. The eTIMSS 2019 Data Collection instruments included both newly developed items selected from the field test and trend items carried forward from TIMSS 2015.

The response process validity of the mathematics PSIs was evaluated based on several sources of qualitative and quantitative data collected during and after the eTIMSS

2019 Field Test. The field test instruments included items that were not selected for main data collection, but because all of the field test items were administered together, these analyses focused on the eTIMSS testing experience and item types without differentiating between items that did and did not move forward after the field test.

The item response data and timing data from the eTIMSS 2019 Field Test was used to analyze the measurement properties and internal structure of the eTIMSS mathematics items. For all item-level analysis, only items that were selected for main data collection were used. The items that are discarded or substantially revised after the field test typically have less desirable measurement properties (e.g., low item difficulty or discrimination, an attractive distracter), content related issues, or are not needed to cover the assessment framework, so excluding these items provided a more accurate representation of the eTIMSS 2019 mathematics assessments. The following sections present the methods used to address the three major research areas.

Validity Evidence Based on Test Content

- Did the methods used to develop the PSIs support a high-quality framework and coherent assessment instruments that minimize construct-irrelevant variance?
- Do the mathematics PSIs address the *TIMSS 2019 Mathematics Framework* and improve coverage of mathematics applying and reasoning skills?

Establishing the content validity of an assessment is primarily achieved through adhering to best practices in assessment design throughout the development process—clearly defining the target construct, specifying the items needed to measure it, and establishing standards for items and test forms to minimize construct-irrelevant variance. Therefore, documentation of the methods and procedures TIMSS used to certify that the

mathematics instruments provide valid measurement of mathematics ability is the first step in establishing the content validity of the eTIMSS 2019 mathematics PSIs.

As described earlier in this chapter, the methods and procedures used for updating the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017) were based on several sources of data provided by the participating countries as well as reviews by international experts to ensure that the target mathematics construct detailed in the framework reflected the goals of the international mathematics education community and curricula of the participating countries. This iterative framework development process also ensured that the frameworks provided clarity for item writers and well-defined specifications for the composition of the assessment. As also described, developing the mathematics PSIs involved an iterative and collaborative effort involving international mathematics experts, representatives from the participating countries, expert item writers, and a good deal of quality control between those developing the PSIs and the programmers. There were numerous reviews to ensure that PSIs reflected the target construct articulated in the framework.

During test development, both the regular and PSI items were classified to the most detailed level of description in the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017). The content and cognitive domain classifications for all items were meticulously reviewed by mathematics and measurement experts to certify that the items are suitable for addressing the mathematics abilities as described in the framework and allow for the item classifications to be used as evidence of content validity. In addition to defining the domains, the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017)

specifies the target percentage of testing time allocated to each content and cognitive domain at the fourth and eighth grade to ensure that the TIMSS 2019 mathematics assessments provided appropriate coverage of mathematics ability at the fourth and eighth grades. When selecting the new regular mathematics items for data collection, both the measurement properties of the individual items and the overall content and cognitive domain coverage of the group of items were considered.

The mathematics PSIs were primarily designed to increase coverage of traditionally difficult to measure areas of the mathematics framework, especially in the applying and reasoning domains. Given these distinct development goals, the tasks were not subject to the same specifications for domain coverage as the regular mathematics items. When developing the PSIs, choices about the mathematics content topics to assess with each task were largely guided by the problem contexts and potential uses of technology to enhance measurement. Still, if the PSIs were to be included in the TIMSS 2019 achievement scale, it was important to confirm that they serve the intended purpose of expanding coverage of the mathematics applying and reasoning cognitive domains, but do not substantially alter the percentage of testing time allocated to each content domain.

To evaluate the extent to which the mathematics PSIs meet these goals, the content and cognitive domain coverage of the regular eTIMSS mathematics items alone was compared to the content and cognitive domain coverage of the full eTIMSS 2019 mathematics assessments with the PSIs. The achieved percentages of score points in each domain with and without the PSIs were compared to the specifications for the target

percentages of testing time allocated to each domain in the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017).

For the content domains, the impact of the PSIs was judged based on the consistency of the content domain coverage with and without the PSIs, as the PSIs were not intended to increase coverage of specific content domains. To provide further detail about the mathematics PSIs and demonstrate their alignment to the framework, a brief description of the problem scenario and the content domain topics addressed in each task also was documented. For the cognitive domains, the impact of the tasks was evaluated based on the change in coverage of the applying and reasoning domains that results from adding the PSIs to the assessments.

In evaluating the impact of the PSIs on the framework coverage of the assessments it is important to keep in mind that the new regular and PSI items developed for eTIMSS 2019 only comprise half the eTIMSS 2019 fourth and eighth grade mathematics assessments. The PSIs constitute approximately 12 percent of the full mathematics assessment at each grade, so including the tasks only was expected to result in minor fluctuations in the overall domain coverage of the assessments.

Validity Evidence Based on Response Process

- Did the user interface, directions, and tools promote ease of navigation and consistency across the tasks?
- Are students' interactions with the eTIMSS mathematics instruments consistent with the cognitive processes the instruments were designed to elicit?
- Can the items that comprise the mathematics PSIs be scored reliably?

As documented earlier in this chapter, data from cognitive interviews, pilot tests, and an ambitious field test led to a series of improvements in the user interface, directions, and tools associated with the eTIMSS assessment, including the PSIs. Several sources of evidence were collected during and after the eTIMSS 2019 Field Test that could be used to evaluate the extent to which students' interactions with eTIMSS and the PSIs were consistent with the cognitive processes the instruments were designed to elicit.

As described in the methods for the eTIMSS 2019 Field Test, students' reactions to the assessment were collected via the eTIMSS Student Questionnaire. Also, timing data indicating the number of seconds students spent on each screen in the eTIMSS 2019 Field Test were captured, allowing for deeper investigation of students' interactions with the items. Following the field test, the NRCs were asked for feedback on the PSIs as well as the eTIMSS operations and procedures via the Field Test Survey Activities Questionnaire. The items from the Field Test Survey Activities Questionnaire that were considered for this dissertation are provided in Appendix B.

In addition to these efforts, the author of this dissertation developed a series of research questions about students' and test administrators' interactions with and reactions to the eTIMSS testing experience, then observed several field test testing sessions in the greater Boston, Massachusetts area in March 2018 to obtain further insight into how students and test administrators interact with eTIMSS, and specifically the PSIs. The questions addressed the usability of the user interface and enhanced item types, student engagement, and the extent to which the test administrator manuals and eTIMSS assessment systems supported test administrators in carrying out the assessment as

intended. A total of four testing sessions (two at each grade) at two different schools were observed. The author summarized the results in a report for staff at the TIMSS & PIRLS International Study Center in April 2018.

Using these sources of evidence, the extent to which the eTIMSS mathematics instruments elicited the intended interactions from students was evaluated based on three criteria—1) the functionality and usability of the eTIMSS interface, tools, item types, and directions, 2) the extent to which students were engaged and motivated by the assessment, and 3) the relationship between the cognitive domain classifications of the items and the amount of time students spent on task. Because the response process validity of the PSIs hinges on the eTIMSS assessment systems, these analyses addressed both the eTIMSS assessment systems in general and aspects of response process validity specific to the PSIs.

For items to be considered valid, it is also essential that they can be scored reliably. The scoring reliability was addressed by reporting the results of the machine and human scoring activities for all of the constructed response items in the mathematics PSIs in the eTIMSS 2019 Field Test, including those that were not selected to move forward to eTIMSS 2019 Data Collection. All of the field test PSIs were used for this analysis so that the full variety of unique item types considered for the mathematics PSIs could be addressed.

For machine-scored items, this included documentation of the number of PSI items that were successfully scored by IEA Hamburg and verified by the TIMSS & PIRLS International Study Center. For human-scored items, all participating countries

were required to double-blind score 100 student responses per item to allow for the percent agreement between the two scorers in terms of both the total score points assigned to the responses (score reliability) and by the specific code (code reliability) to be evaluated. The results of the within-country reliability scoring activities were used to evaluate the extent to which the scoring guides and scorer training were successful in supporting consistent application of the scoring guides.

Validity Evidence Based on Internal Structure

- Do the properties of the mathematics items that comprise the PSIs differ from the regular eTIMSS mathematics items? And if so, how?
- How do the PSIs fit with the hypothesized factor structure underlying mathematics ability?

The timing data and students' responses to the mathematics items from the eTIMSS 2019 Field Test were used to begin to evaluate the measurement properties and internal structure of the eTIMSS 2019 mathematics assessments. Because only the data from the field test were available, these analyses are considered to be a preliminary examination of the internal validity of the assessments that will be further investigated with the eTIMSS 2019 data once data collection is completed.

First, the timing data were used to detect evidence of speededness (i.e., students spending less time on items because of insufficient testing time) and position effects (i.e., differences in the amount of time students spent on items depending on their order of appearance in the assessment), which could have impacted the measurement properties of the items in the field test. Then, the item response data from the eTIMSS 2019 Field Test were used to compare the measurement properties of the PSI items to the regular

mathematics items and investigate the consistency of countries' performance across the two item types. Finally, the underlying factor structure of the regular mathematics items and items within the mathematics PSIs was evaluated by fitting a series of factor analysis models to the student response data and comparing the fit of these models.

Speededness and Position Effects

The total testing time for all eTIMSS block combinations was 72 minutes at the fourth grade and 90 minutes at the eighth grade. With a total of four item blocks in each test form (two mathematics and two science), each fourth grade block was designed to comprise approximately 18 minutes of testing time and each eighth grade block was designed to comprise approximately 22.5 minutes of testing time. Based on previous paper-based assessments, TIMSS has established standards for the number of regular mathematics items in each fourth and eighth grade block. TIMSS blocks typically contain 10 to 13 items and 12 to 16 score points at the fourth grade and 13 to 16 items and 14 to 18 score points at the eighth grade.

The blocks in the eTIMSS 2019 Field Test were designed to follow these conventions, but introducing a new mode of administration, enhanced item types, and interactive feature has the potential to impact the number of items students can reasonably complete in the given testing time, particularly for the PSIs. In addition to including more enhanced features than the regular blocks, the PSI blocks generally included more score points because they included more complex applying and reasoning items. The TIMSS assessments are not designed to be speeded tests, so it is important to

confirm that the positionality of the items within the instruments and timing restrictions did not adversely influence students' responses.

First, the international average amount of time students spent on each regular block and PSI block was calculated. Under the field test block combination design, each mathematics PSI block appeared in two block combinations—once as the first block in the mathematics part of the instrument and once as the second block. For the PSI blocks, the average total time per screen was calculated separately for the two positions in which the block appeared.

To detect evidence of speededness, the average total time for each mathematics block was compared to the amount of testing time allocated to a block under the assessment design. Then, to investigate position effects, the average time for each PSI block was compared across the two positions in which it appeared. The average amount of time for the PSI blocks was also compared to the regular mathematics blocks to identify possible differences in the average time students spent on the two block types.

Measurement Properties of the Items

The item response data from the eTIMSS 2019 Field Test was then used to evaluate the measurement properties of the mathematics items that were selected for main data collection. The international average item difficulty, item discrimination, percent omitted, and percent not reached were compared across the PSI items and regular eTIMSS mathematics items to investigate differences in the measurement properties of the two item types.

Once the eTIMSS 2019 achievement instruments were finalized, the field test item statistics were recalculated with only the items that were selected for data collection to allow for reasonable comparisons to be made across the regular and PSI items. Exhibit 3.9 shows the number of valid items and responses from the eTIMSS 2019 Field Test that were used in this analysis, as well as the evaluations of the performance consistency across regular and PSI items and investigation of the underlying factor structure.

Exhibit 3.9: Number of Items and Responses from the eTIMSS 2019 Field Test Used in Analysis

Grade	Total Cases	Regular Items		PSI Items	
		Valid Items	Average Responses per Item*	Valid Items	Average Responses per Item*
Grade 4 (31 countries)	43,293	76	6,607	27	6,488
Grade 8 (22 countries)	31,116	86	4,748	22	4,700

*Counts reflect resulting sample sizes after deleting problematic data.

Performance Consistency across Regular and PSI Items

Each countries' average percent correct was calculated separately for the regular items and PSI items to investigate the extent to which countries' achievement was consistent across the two item types. Scatter plots of countries' average percent correct on the PSI items against their average percent correct on the regular items and the correlation between these two percent correct scores were used to evaluate strength and consistency of the relationship between countries' performance on the two item types. Consistent with the methods used in evaluating the measurement properties of the items, only the items that were selected for data collection were used to calculate countries' average percent correct.

Underlying Factor Structure

The underlying factor structure of the regular eTIMSS mathematics items and items within the mathematics PSIs was evaluated by fitting a series of confirmatory factor analysis models to the student response data from the eTIMSS 2019 Field Test and comparing the fit of these models to the data. A unidimensional model, two-dimensional model, and bi-factor model were used. Again, only the items from the field test that were selected for data collection were included in these analyses. All models were estimated in Mplus Version 8 (Muthén & Muthén, 1998–2017).

A confirmatory approach was chosen over an exploratory approach because the purpose of this analysis was to investigate the *a priori* theory that both the regular and PSI mathematics items measure students' mathematics ability (i.e., are a unidimensional construct). Comparing the fit of several competing models provided evidence of the extent to which the items draw upon students' mathematics abilities as intended and helped to identify differences between the regular and PSI items.

The characteristics of the eTIMSS 2019 Field Test data presented challenges in analyzing the factor structure of the assessments, which in turn guided the methods used in these analyses. The following sections describe the characteristics of the data and analysis approaches used to overcome these challenges. Then, the series of models fit to the data are presented, followed by the criteria used to select the preferred model.

eTIMSS 2019 Field Test Data

The data files used for these analyses were obtained in SPSS Statistics Software Version 24 (IBM Corp., 2016) format from the data analysis team at the TIMSS & PIRLS International Study Center. For multiple-choice items, the data files included the

response options selected by students (e.g., A, B, C, D). For constructed response items, the data files included the two-digit score codes assigned to students' responses based on the unique scoring guide developed for each item (see section on *Developing Scoring Guides for Constructed Response Items*). A SPSS program provided with the data files was used to assign score levels, or point values, to the raw response data according to the answer keys for multiple-choice items and scoring guides for constructed response items in preparation for analysis. The eTIMSS assessments at both grades included both dichotomous items worth one score point (scored as 1=correct, 0=incorrect) and polytomous items worth two score points (scored as 2=fully correct, 1=partially correct, 0=correct). All missing responses were recoded to 9.

For field test data collection, the pool of items for each subject and grade was grouped into 13 item blocks—10 blocks comprised of regular eTIMSS items and three blocks comprised of PSIs. At each grade, the blocks were arranged in a total of eight unique block combinations, each including two blocks of mathematics items and two blocks of science items (see Exhibit 3.5). The 10 regular blocks were distributed across five of these block combinations using an incomplete and un-rotated design in which each block only appeared in a single block combination. The three PSI blocks were distributed across the remaining three block combinations using a balanced incomplete block design, in which each PSI block appeared twice—once with each of the other PSI blocks—and were rotated across the block combinations.

Each student participating in the field test completed a single block combination, or approximately 15 percent of the total item pool. Within the participating classes, the

block combinations were distributed among students according to predetermined random assignments produced by the TIMSS within-school sampling software. This means that the sample of students completing each block combination in each country was approximately randomly equivalent in terms of student ability (Martin, Mullis & Foy, 2017).

The un-rotated incomplete design used for the regular block combinations enabled TIMSS to try out as many regular items as possible in the field test, but presented challenges in analyzing student responses to all of the mathematics items together. To fit a confirmatory factor analysis or item response theory model to the data, an inter-item covariance matrix of the observed relationships among all pairs of items (i.e., each mathematics item with every other mathematics item) is needed. As a result of the block combination design, the student-level mathematics data could not be used to produce a complete covariance matrix.

Exhibit 3.10 shows the data matrix of responses from the eTIMSS 2019 Field Test. For each mathematics and science block, the shaded cells show the block combinations in which the block appeared, or where there were student responses to the items. The pairs of regular items that appeared in different block combinations (e.g., an item in ME01 and an item in ME03) were never completed by a common group of students. Therefore, it was not possible to establish the covariance among these pairs of items. In other words, because each regular block only appeared in one block combination in the field test, there was no mechanism for linking student responses across the regular block combinations. Also, because no regular and PSI items appeared

together in a block combination, there was no mechanism for linking student responses to the regular and PSI items either. Without these links, the mathematics data alone could not be used in its original format to test the underlying factor structure of the assessment.

Exhibit 3.10: eTIMSS 2019 Field Test Data Matrix

Block		Regular Block Combinations					PSI Block Combinations		
		1	2	3	4	5	6	7	8
Mathematics Blocks	ME01								
	ME02								
	ME03								
	ME04								
	ME05								
	ME06								
	ME07								
	ME08								
	ME09								
	ME10								
	MI01								
	MI02								
	MI03								
Science Blocks	SE01								
	SE02								
	SE03								
	SE04								
	SE05								
	SE06								
	SE07								
	SE08								
	SE09								
	SE10								
	SI01								
	SI02								
	SI03								

Blocks beginning with “ME” and “SE” are regular eTIMSS mathematics and science blocks, respectively. Blocks beginning with “MI” and “SI” are mathematics PSI blocks and science PSI blocks, respectively.

Each student participating in the eTIMSS 2019 Field Test completed one block combination. Shaded cells indicate where there are student responses to items.

The structure of the field test data was further complicated by the two-stage random sampling design TIMSS used to select the sample of schools, and then intact classes of students within the schools, to participate in the assessment. Randomly

selecting schools and classes of students rather than individual students introduced clustering into the data because students in the same class are more likely to have similar responses to items. Therefore, the observed student responses in the eTIMSS 2019 Field Test dataset cannot be considered completely independent, which must also be taken into account in analysis to avoid violating a fundamental assumption of factor analysis and item response theory models.

Analysis Approaches

Two different analytic approaches were used to overcome the issue of insufficient links across items and respondents—1) aggregating the mathematics item responses to the class level to produce a complete covariance matrix based on class means, and 2) adding students' responses to the science items to the dataset to provide more common links across the block combinations (e.g., science PSI block SI01 appears in both block combination 6 and block combination 8). Under each approach, a different method was used to address the dependencies among the observed responses. Both approaches have known benefits and disadvantages, but considering the results of two approaches can help to strengthen the conclusions made about the structure of the data.

Approach 1: Class-level analysis

First, students' responses to the mathematics items were aggregated to the class level, such that the observed responses for each item became the mean score of all students in a class that received an item. With the random sampling methods used to assign the block combinations to students, the students across each country completing each block combination are known to be approximately equivalent in terms of achievement and some students in every sampled class completed each block

combination. Therefore, using the class-level data, a complete covariance matrix could be established and used to test all models of interest.

However, this approach has some limitations. Models estimated with the aggregated data analyze the differences in responses between groups of students in different classrooms (the between-class variance) rather than the variance among individual students (the student within-class variance) captured in the original student-level data. In most TIMSS countries, the student within-class variance accounts for the majority of the total variance in students' mathematics achievement (Martin, Foy, Mullis & O'Dwyer, 2013; Gustafsson, Nilsen & Hansen, 2018). The amount of the total variance in student achievement accounted for at the class level has been shown to vary widely across TIMSS countries, ranging from 5 percent to 67 percent of the total variance, and is commonly related to class-level contextual variables such as classroom resources and the teaching methods that do not show as strong an effect at the student-level (Martin et al., 2013; Gustafsson et al., 2018). Therefore, analyzing the class-level data can provide a similar view of the underlying factor structure as the student-level data, but the results from this approach must be carefully interpreted with this shortcoming in mind.

Aggregating the data to the class-level also substantially reduced the sample size. Nevertheless, given the size of the original dataset, the dataset used with this approach was still relatively large (Grade 4: $n = 2,163$ classes; Grade 8: $n = 1,403$ classes) so the loss in sample size was a less impactful limitation.

Approach 2: Student-level analysis including science items

Although the student-level mathematics data could not be analyzed alone due to the lack of overlapping items across block combinations, it was possible to conduct some student-level analyses when students' responses to the science items in the eTIMSS 2019 Field Test were included in the models. Given that mathematics and science performance are correlated in the student population, students' performance on the science items can help estimate mathematics ability, and therefore the mathematics item parameters.

As shown in Exhibit 3.10, the science PSI blocks were assigned to the block combinations using the same balanced incomplete block design as the mathematics PSI blocks. Therefore, including the science PSI items in these analyses increased the number of items in the dataset that appeared in more than one block combination, which allowed for the covariances among more pairs of items to be established. Also, although the regular science blocks did not provide any links across block combinations, students' responses to these items provided more information about their ability on a correlated construct, which helped to meet the criteria for model convergence. Although there was still a substantial amount of planned missing data after the science items were added to the dataset, the additional links across block combinations and observed responses to items from each student provided enough information about students' abilities to fit some of the models of interest with the student-level data.

However, introducing another construct into the models (science ability) may have a small impact on the model goodness-of-fit indices and the mathematics item parameters. This is because students' responses to the science items will contribute to the

determinant of the input matrix (the scalar that reflects a generalized measure of variance for the entire set of variables in the matrix), which is used to estimate all parameters in the model (Brown, 2014). Still, because students' mathematics and science ability are known to be correlated and the science items comprise a separate factor in these models, the impact of adding these items to the model is expected to be minimal.

When analyzing the student-level data, the nonindependence of the observed responses due to the clustering of students within classes also must be addressed. For these analyses, a design-based approach was used in Mplus to account for the variance structure resulting from the complex two-stage clustered sampling design used in the field test by including sampling and weighting variables in the models to define how the sample was drawn from the target population (Muthén & Satorra, 1995; Wu & Kwok, 2012). The models fit with this approach were estimated with the sampling zone (JKZONE) and sampling replicate (JKREP) weighting variables used to define strata and clusters (Rust, 2014). The cases were weighted using the TIMSS senate weight variable (SENWGT), which gives equal weight to each country in the analysis (Foy, 2017; LaRoche, Joncas & Foy, 2017).

This design-based approach adjusts the parameters estimates and standard errors for the model based on the sampling design and is commonly used when the primary purpose of the analysis is to validate the student-level covariance structure of an assessment (Wu & Kwok, 2012; Muthén & Asparahov, 2006). When the underlying factor structure at the student-level and class-level are the same, this approach has been shown to perform equally as well as more complex multilevel models (i.e., model-based

approaches) that separately model the class- and student-level factor structure (Muthén & Satorra, 1995). Because the purpose of these analyses was to investigate the items' relationships with students' mathematics ability, the simpler design-based approach was chosen. Nevertheless, defining how the sample was drawn from the target population does not directly model the class-level variance structure, so the results from these analyses also should still be interpreted with this limitation in mind.

Factor Analysis and Item Response Theory Models

Several confirmatory factor analysis (CFA) models and item response theory (IRT) models were used in these analyses. For the models fit with the class-level data, linear CFA models were used because the aggregated data are continuous. For the student-level data including the science items, IRT models allowed for modelling the non-linear relationship between the categorical observed variables and latent constructs. The following sections present the CFA models, then the parallel IRT models. Additional details about the theories and intended uses of these models were provided in Chapter 2 (section on *Validity Based on Internal Structure*).

Confirmatory Factor Analysis

In the traditional CFA model (Jöreskog, 1969; 1971a) each item is assigned to one of a number of factors that each represent a latent variable. The response of person i on item Y assigned to factor j , Y_{ij} can be expressed as:

$$Y_{ij} = u_{ij} + \lambda_{ij}\eta_j + \varepsilon_{ij} \quad (1)$$

where u_{ij} is the intercept of Y_{ij} ; λ_{ij} is the factor loading of Y_{ij} on η_j ; η_j is the factor for the group of items; and ε_{ij} is the specific error for Y_{ij} . The factor loadings are linear regression slopes for predicting the observed responses from the latent variable. It is assumed that the factors and errors are independent (η_j is not correlated with ε_{ij}), the errors are independent of each other ($\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$), the factors have a mean of zero and variance of one, and the errors have a mean of zero and their specific variances.

For the unknown parameters (freely estimated elements) of a CFA model to be identified, the number of parameters must be less than the number of entries on and below the diagonal of the observed covariance matrix. The degrees of freedom for a model is the difference between these two quantities.

Confirmatory Bi-factor Model

The bi-factor model (Gibbons & Hedeker, 1992) is an extension of the CFA model that includes both a general factor for the overarching latent variable the instrument is designed to measure as well as specific factors for groups of items that are expected to share unique common variance beyond the general factor. All items are assigned to both the general factor and one specific group factor. This extended CFA model with k specific factors can be expressed as:

$$Y_{ij} = u_{ij} + \lambda_{G,ij}\eta_G + \lambda_{S1,ij}\eta_{S1} + \dots + \lambda_{Sk,ij}\eta_{Sk} + \varepsilon_{ij} \quad (2)$$

where u_{ij} is the intercept of Y_{ij} ; $\lambda_{G,ij}$ is the factor loading on the general factor; η_G is the general factor measured by all items; $\lambda_{S,ij}$ is a factor loading of an item on its specific factor; η_S are the specific factors; and ε_{ij} is the specific error for Y_{ij} .

In addition to the assumptions of the CFA model, the correlations among the specific factors are fixed to zero and the correlations between the general factor and each specific factor is fixed to zero to identify the model (Rijmen, 2011). The mean and variance of each factor are set to zero and one, respectively, so it can be assumed that the latent variables follow a standard normal distribution (Rijmen, 2009).

Item Response Theory Models

Under the IRT framework, a probability in the logit metric is used to model the relationship between individuals' responses to items and scores on latent factors. The linear equations for the CFA and confirmatory bi-factor models (Equations 1 and 2, respectively) are re-expressed as the probability of an individual responding correctly to an item on an instrument based on their latent trait score. When each item is only assigned to one latent trait (analogous to Equation 1), the probability of an individual scoring in the l^{th} category (z_{il}) is expressed as:

$$z_{il} = \alpha_i \theta + c_{il} \quad (3)$$

where α_i is the discrimination of an item on a factor; θ is an individual's latent trait score; and c_{il} is a multidimensional intercept parameter equal to the negative product of the factor loading and the threshold parameter for the l^{th} category.

In the IRT bi-factor model (analogous to Equation 2), the probability of an individual scoring in the l^{th} category (z_{il}) is determined by an individual's latent trait score on the general factor and the k specific factors in the model. The IRT bi-factor model can be expressed as:

$$z_{il} = \alpha_{iG}\theta_G + \alpha_{iS1}\theta_{S1} + \dots + \alpha_{iSk}\theta_{Sk} + c_{il} \quad (4)$$

where α_{iG} is the discrimination of an item on the general factor; θ_G is an individual's latent trait score on the general factor; α_{iS} is the discrimination of an item on its specific factor; θ_S is an individual's latent trait score on a specific factor; and c_{il} is a multidimensional intercept parameter equal to the negative product of the factor loading and the threshold parameter for the l^{th} category. Consistent with the linear version of the model, the correlations among the specific factors and correlations between the general factor and all specific factors are fixed to zero and the mean and variance of each dimension is set to zero and one, respectively.

Different link functions can be used to express the probability of a correct response (z_{il}) for binary (1-point) and polytomous (2-point) items. For binary items in these analyses, Mplus used the two-parameter logistic model (Birnbbaum, 1968) to calculate the probability of a correct response:

$$z_{il} = p_{(x_j=1|\theta, \delta_j, \alpha_j)} = \frac{e^{\alpha_j(\theta - \delta_j)}}{1 + e^{\alpha_j(\theta - \delta_j)}} \quad (5)$$

where j is an item; θ is the persons' latent trait score; δ_j is item j 's difficulty; and α_j is item j 's discrimination parameter.

For polytomous items, Mplus used the graded response model (Samejima, 1969):

$$z_{il} = P_{x_j}^*(\theta) = \frac{e^{\alpha_j(\theta - \delta_{xj})}}{1 + e^{\alpha_j(\theta - \delta_{xj})}} \quad (6)$$

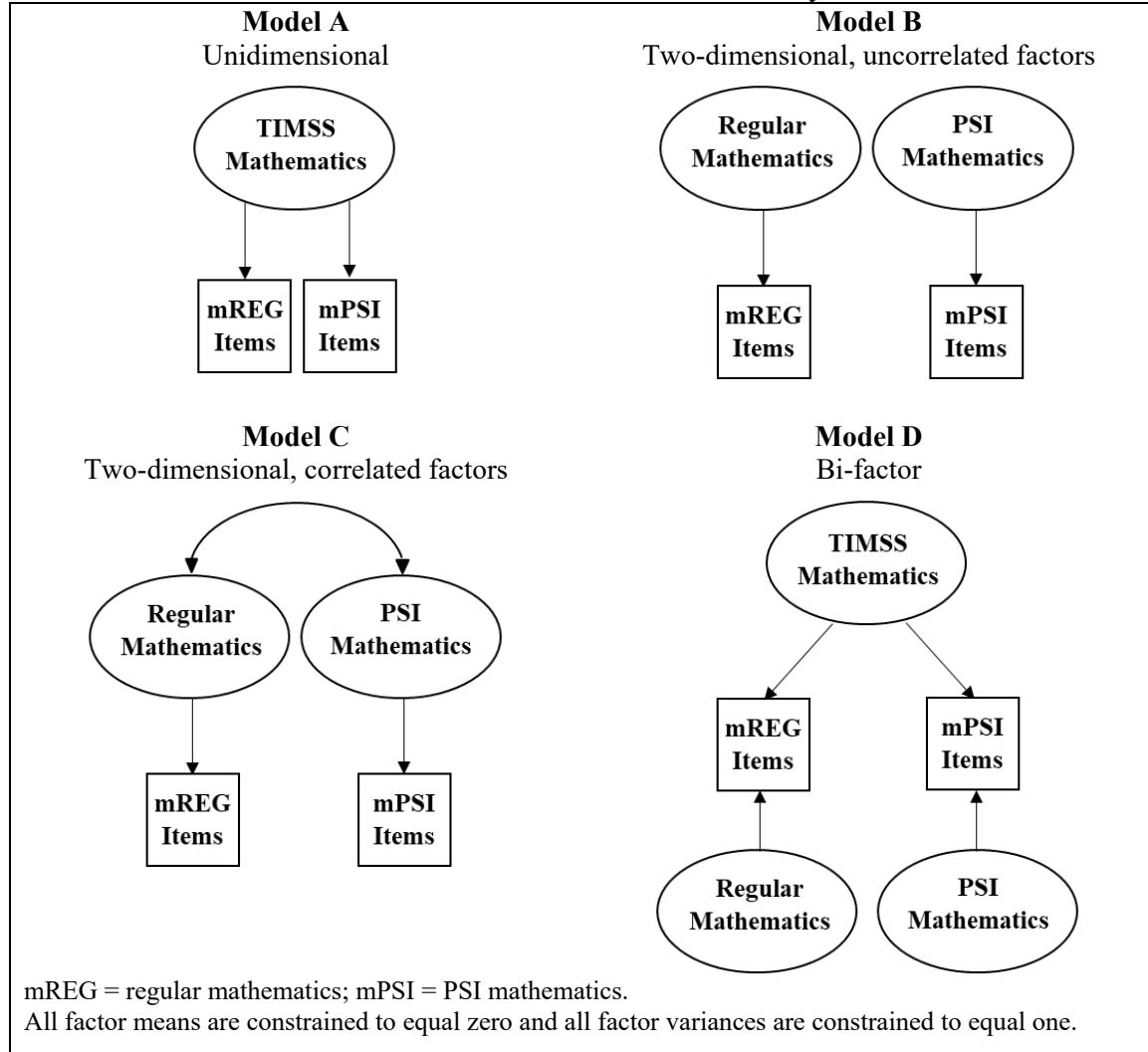
where j is an item; $P_{x_j}^*(\theta)$ is the probability of responding category x_j or higher to item j ; θ is the persons' latent trait score; δ_{xj} is the category boundary for x_j , or the boundary between categories k and $k-1$; and α_j is item j 's discrimination parameter.

Analysis Models

Class-level analysis

Exhibit 3.11 presents the series of CFA and bi-factor models fitted to the class-level data. In the diagrams, the ovals represent the factors and the squares represent the vectors of responses to the groups of items loading on each factor. The single-headed arrows pointing from the factors to the observed item responses indicate that the latent variables are viewed as the cause of the observed item responses. The curved double-headed arrow between factors in Model C indicates that the factors are allowed to freely correlate.

Exhibit 3.11: Analysis Models Used to Investigate the Underlying Structure of the eTIMSS 2019 Mathematics Field Test Data – Class-level Analysis



Model A was a unidimensional model with a single latent factor for both the regular and PSI items. This model assumes that all the mathematics items measured the same mathematics ability and that there are no meaningful differences between the item types. Model A was considered the baseline model to which all other competing models were compared because both the regular and PSI items were designed to measure the same construct and the most parsimonious model that fits the data is preferred. Models B

and C were two-dimensional models, with separate correlated factors for the regular and PSI items. In Model B the correlation between the two factors was fixed at zero, treating the latent variables as completely unrelated constructs, while in Model C the regular and PSI factors were allowed to correlate freely, so that the relationship between the separate regular and PSI factors could be investigated. Model D was a bi-factor model with a general factor for mathematics and specific factors for the item types. In this model, all of the items are regarded as measures of the same mathematics ability like in the unidimensional model, but any shared residual variance among the groups of items beyond the general factor was also modeled with the specific factors.

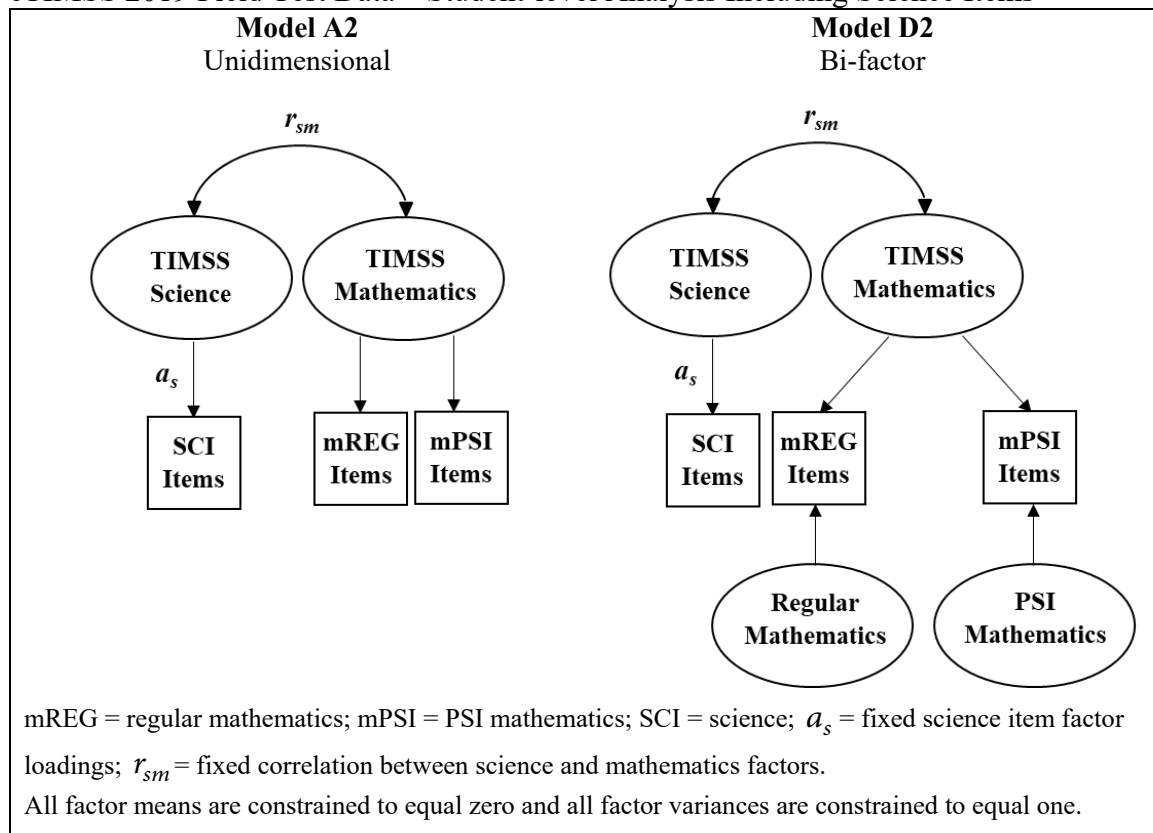
For all models, the mean and variance of each factor were fixed to zero and one, respectively, so it could be assumed that the latent variables follow a standard normal distribution. In Model D, the correlations between each pair of factors (general and specific) were fixed to zero to identify the model.

The models were estimated using maximum likelihood (ML) estimation. ML is an iterative estimation process that aims to find the parameter estimates for the model that maximize the likelihood of these parameters given the observed data (Brown, 2014). Using an initial set of starting values for the parameter estimates, ML repeatedly refines the estimates until arriving at a set that cannot be further improved upon to reduce the difference between the predicted covariance matrix and the sample covariance matrix (i.e., model convergence) (Brown, 2014).

Student-level analysis including science items

The unidimensional model and bi-factor model were fit a second time using the student-level data including the science items (herein referred to as Model A2 and D2, respectively). Due to the lack of overlap across regular and PSI item blocks, the two-dimensional models with separate factors for regular and PSI items (Models B and C) could not be fit with the student-level data. For Models A2 and D2, IRT link functions were used because the data are categorical. These models are shown in Exhibit 3.12.

Exhibit 3.12: Analysis Models Used to Investigate the Underlying Structure of the eTIMSS 2019 Field Test Data – Student-level Analysis Including Science Items



Consistent with the method used for the mathematics items, only the science items that were selected for eTIMSS 2019 Data Collection were included in these analyses. At both grades, a preliminary two-dimensional IRT model with factors for

mathematics and science was fit to obtain estimates for the science item parameters and the correlation between mathematics and science scores. These values were then fixed in subsequent models to reduce the computational demands of fitting the more complex bi-factor model and minimize the impact of the science items on estimates of the mathematics item parameters.

The student-level models were fit with maximum likelihood estimation with robust standard errors (MLR in Mplus) instead of the traditional ML estimation used with the continuous class-level data. Standard ML estimation relies on the assumption that the observed data are normally distributed, so a variant of this technique that allows for corrections to be made to account for violations of this assumption was needed for the models fit with the categorical student-level data. Using standard ML estimation when the normality assumption is not met can have a variety of detrimental consequences, including reduced precision and accuracy of the parameter estimates, spuriously inflated significance tests, and biased factor loadings and standard errors (Kaplan, 2009; Muthén & Kaplan, 1985, 1992).

MLR was selected because it allows for violations of the normality assumption with ordinal data and has been identified as the best approach for analyses with a large number of variables, observations, and missing data, all of which are applicable to the field test data (Brown, 2014; Muthén, Kaplan, Hollis, 1987; Muthén, Muthén & Asparouhov, 2015). MLR can also be used to correct for the impact of the sample design, including stratification, non-independence of observations, and unequal probability of selection (Muthén & Muthén, 1998–2017).

The missing data in the dataset were treated as missing at random (MAR) because any differences between the missing and non-missing cases can be entirely explained by the block combination assignments (Muthén et al., 1987). Because the data are MAR, pairwise deletion (i.e., only deleting cases from correlations in which one or both of the items were not answered) could be used to handle the missing data in these analyses (Muthén & Muthén, 1998–2017). This allowed for all available responses to each item to be retained.

Criteria for Evaluating Model Fit

For both analysis approaches, the relative fit of the series of models was compared based on the Akaike Information Criterion (AIC; Akaike, 1973), Bayesian Information Criteria (BIC; Schwarz, 1978), and the standardized factor loadings. The AIC and BIC were selected for use in these analyses over absolute fit indices (e.g., chi-square tests) because absolute indices become inflated with large sample sizes and can lead to false conclusions about statistical significance of differences between models (Brown, 2014).

The AIC and BIC are parsimony correction model fit indices calculated using the estimated $-2\log\text{likelihood}$ ($-2LL$), which provides an indication of how well the model fit the data, and “penalties” for other characteristics of the model that influence the observed fit (Brown, 2014). Models with more parameters naturally provide better fit, so these parsimony correction indices are useful in identifying the simplest and best fitting model for the data (Brown, 2014). The AIC includes a penalty for the number of freely estimated parameters in the model. The BIC includes penalties for both the number of the

freely estimated parameters and the sample size. The AIC tends to favor more complex models when the sample size is large, while the BIC remains relatively stable across different sample sizes, so both were considered (DeMars, 2013).

Equations 7 and 8, respectively, provide the equations for calculating the AIC and BIC, where b is the number of freely estimated parameters in the model and N is the sample size:

$$AIC = -2LL + 2b \quad (7)$$

$$BIC = -2LL + b(\ln(N)) \quad (8)$$

The absolute value of the AIC and BIC alone are meaningless, but the relative AIC and BIC values of a series of competing models fit to the same data can be compared to select the best fitting model in the series (Burnham & Anderson, 2002). The model with the lowest AIC and BIC is considered to have the best fit. The change in AIC and BIC between this model and each of the other models (ΔAIC and ΔBIC , respectively) can be compared to determine whether there is enough empirical evidence to support the conclusion that the best fitting model provides substantially better fit than each competing model (Burnham & Anderson, 2002).

The models fit to the data also were evaluated based on the items' standardized factor loadings, which indicate the strength of the relationship between the items and factors. For the analyses using the class-level data, these loadings are the linear regression slopes for predicting the observed responses from the latent variables. For the analysis using the student-level data including the science items, the loadings are the items' discriminations on the factors. In both cases, these values range from -1 to 1 and

higher values are interpreted as evidence of a stronger relationship between the item and the factor.

Using the approach described in Rijmen (2011), the median factor loadings (λ_{median} for linear factor models and α_{median} for IRT models) for all items assigned to each factor were used to evaluate the overall strength of the relationship between the groups of items and the factors. With a large number of items in the assessment, this approach was helpful in interpreting the general relationship between the items and factors.

For the bi-factor models (Models D and D2), scatterplots were used to further investigate the relationship between the standardized factor loadings on the general factor and specific factors. The items were plotted with their general factor loadings on the x -axis and specific factor loadings on the factor to which they were assigned (regular or PSI) on the y -axis. Visual inspection of the clustering of the items on these plots was used to determine whether the residual variance shared among the groups of items was substantial enough to warrant the use of specific factors.

Chapter 4: Results

Chapter 3 described the rigorous and lengthy procedures TIMSS used to develop the PSIs, including adhering to the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017) and conducting numerous expert reviews. As part of ensuring the validity of the test content and response process, the process was supported by cognitive interviews, pilot tests, and an ambitious field test. Chapter 3 also provided information about the field test data and presented a number of analyses used to address the validity of the mathematics PSIs in the context of the full eTIMSS 2019 assessment.

Chapter 4 presents the results of the analyses discussed in Chapter 3. The results of these analyses provide more information about the test content, response process validity, and internal structure validity, including the application of a series of factor analysis models used to examine the structure of the relationships between the PSIs and the regular eTIMSS mathematics items.

Taken together, the information in Chapters 3 and 4 are used to build a coherent validity argument that will support interpretations of scores on the eTIMSS 2019 mathematics PSIs and address the overarching research question: Does adding the PSIs to the eTIMSS mathematics assessments enhance the validity of the TIMSS mathematics achievement scales at the fourth and eighth grades?

The results for the fourth and eighth grade mathematics assessments are discussed together and organized by the three validity research areas covered in this dissertation—test content, response process, and internal structure. The chapter concludes with a summary of the key findings.

Test Content Validity

After a four-year development process and full-scale field test, three mathematics PSIs at each grade were selected for eTIMSS 2019 Data Collection. At both grades, the tasks covered a range of topics in the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017) with items situated in a variety of problem contexts. Exhibits 4.1 and 4.2 provide a brief description of the PSI problem scenarios and the mathematics content domain topics assessed with the fourth and eighth grade mathematics PSIs, demonstrating the tasks' alignment to the framework. The number of score points allocated to each of the content domain topics within each PSI are shown in parentheses after the topics.

Exhibit 4.1: Fourth Grade eTIMSS 2019 Mathematics PSI Problem Scenarios and Framework Content Domain Topics within Number, Measurement and Geometry, and Data

Grade 4 PSIs	Content Domain Topics (Score Points)
<i>School Party</i> Students plan a party for a school by determining the price for tickets and the amount of food, drinks, and decorations to purchase for the party.	<ul style="list-style-type: none"> • Whole Numbers (7) • Expressions, Simple Equations, and Relationships (2) • Fractions and Decimals (1) • Reading, Interpreting, and Representing Data (2) • Using Data to Solve Problems (2)
<i>Robots</i> Students use a robot that can follow input-output rules to solve mathematics problems and determine the robot's rules.	<ul style="list-style-type: none"> • Whole Numbers (2) • Expressions, Simple Equations, and Relationships (5)
<i>Little Penguins</i> Students add information to a website about Little Penguins by solving a series of mathematics problems involving facts about penguins.	<ul style="list-style-type: none"> • Whole Numbers (7) • Expressions, Simple Equations, and Relationships (1) • Measurement (4) • Reading, Interpreting, and Representing Data (2)

() Score points are shown in parentheses.

The fourth grade mathematics PSIs selected for eTIMSS 2019 Data Collection included a total of 29 items, worth a total of 35 score points.

Exhibit 4.2: Eighth Grade eTIMSS 2019 Mathematics PSI Problem Scenarios and Framework Content Domain Topics within Number, Algebra, Geometry, and Data and Probability

Grade 8 PSIs	Content Domain Topics (Score Points)
<i>Dinosaur Speed</i> Students use the relationships between foot length, leg height, and stride length to estimate how fast a dinosaur could run.	<ul style="list-style-type: none"> • Ratio, Proportion, and Percent (1) • Expressions, Operations, and Equations (5) • Relationships and Functions (3) • Geometric Shapes and Measures (2) • Data (1)
<i>Building</i> Students determine the dimensions of a shed, including a barrel to collect rainwater.	<ul style="list-style-type: none"> • Expressions, Operations, and Equations (3) • Geometric Shapes and Measurements (8)
<i>Robots</i> Students determine functions using a robot that uses a function to determine y for any given value of x .	<ul style="list-style-type: none"> • Relationships and Functions (4)

() Score points are shown in parentheses.

The eighth grade mathematics PSIs selected for eTIMSS 2019 Data Collection included a total of 25 items, worth a total of 27 score points.

eTIMSS Mathematics Framework Coverage with the PSIs

This dissertation evaluated the impact of adding the PSIs to the eTIMSS 2019 mathematics assessments at the fourth and eighth grades by comparing the mathematics framework coverage provided by the eTIMSS 2019 mathematics assessments with and without the PSIs. As shown in the next two sections, the detailed analyses of framework coverage across the content and cognitive domains provide evidence that the mathematics PSIs could be included in the TIMSS 2019 achievement scale from a content validity perspective.

Content Domain Coverage

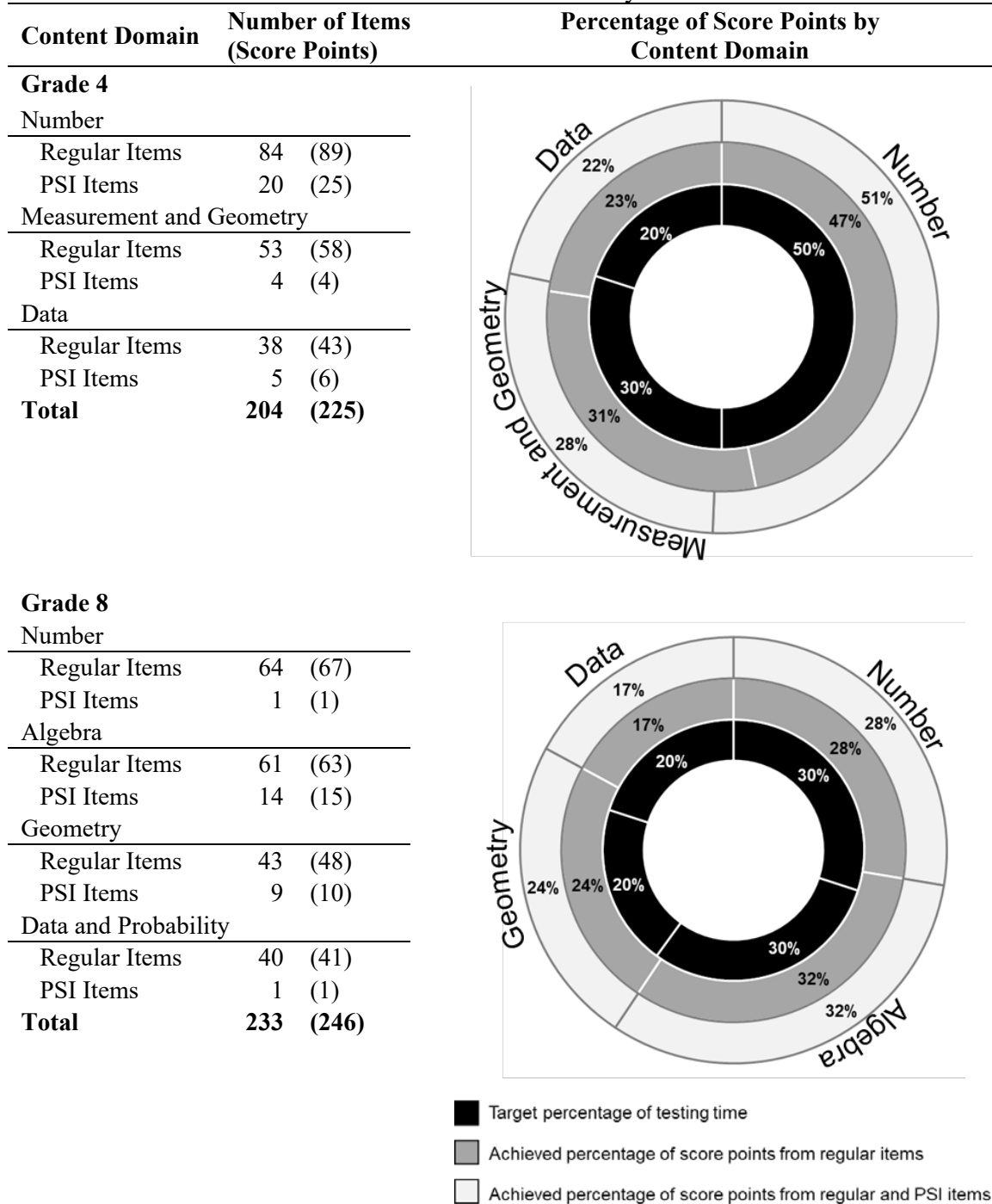
Exhibit 4.3 presents the number of items and score points in the fourth and eighth grade eTIMSS 2019 mathematics assessments by the content domains in the framework

and item type (regular or PSI) together with multi-level pie charts showing the percentage of assessment score points in each content domain. For each grade, the center ring of the pie chart shows the target percentage of testing time allocated to each content domain in the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017), the middle ring shows the achieved percentage of score points for the regular eTIMSS mathematics items, and the outer ring shows the achieved percentage of score points for the full eTIMSS assessment, including both regular and PSI items.

At both grades, the achieved content domain coverage of the regular items was within four percentage points of the target percentages of testing time specified in the framework. Given the many other development objectives for the assessments (e.g., cognitive domain coverage, variety of item formats, item difficulty and discrimination), some deviation from the target percentages of testing time was expected and these results may be considered sufficiently consistent with the framework specifications.

The mathematics PSIs primarily focused on topics in one or two content domains in each grade. Consistent with the content areas emphasized in the mathematics framework, the fourth grade PSIs mainly addressed topics within the number content domain, increasing coverage of this domain by a total of 20 items and 25 score points. At the eighth grade, the PSIs mainly addressed topics in the algebra and geometry content domains, contributing an additional 23 items and 25 score points in these two domains beyond the regular items.

Exhibit 4.3: eTIMSS 2019 Mathematics Assessments by Content Domain



The fourth grade eTIMSS 2019 mathematics assessment included 175 regular items, worth a total of 190 score points, and 29 PSI items, worth a total of 35 score points. The eighth grade eTIMSS 2019 mathematics assessment included 208 regular mathematics items, worth a total of 219 score points, and 25 PSI items, worth a total of 27 score points.

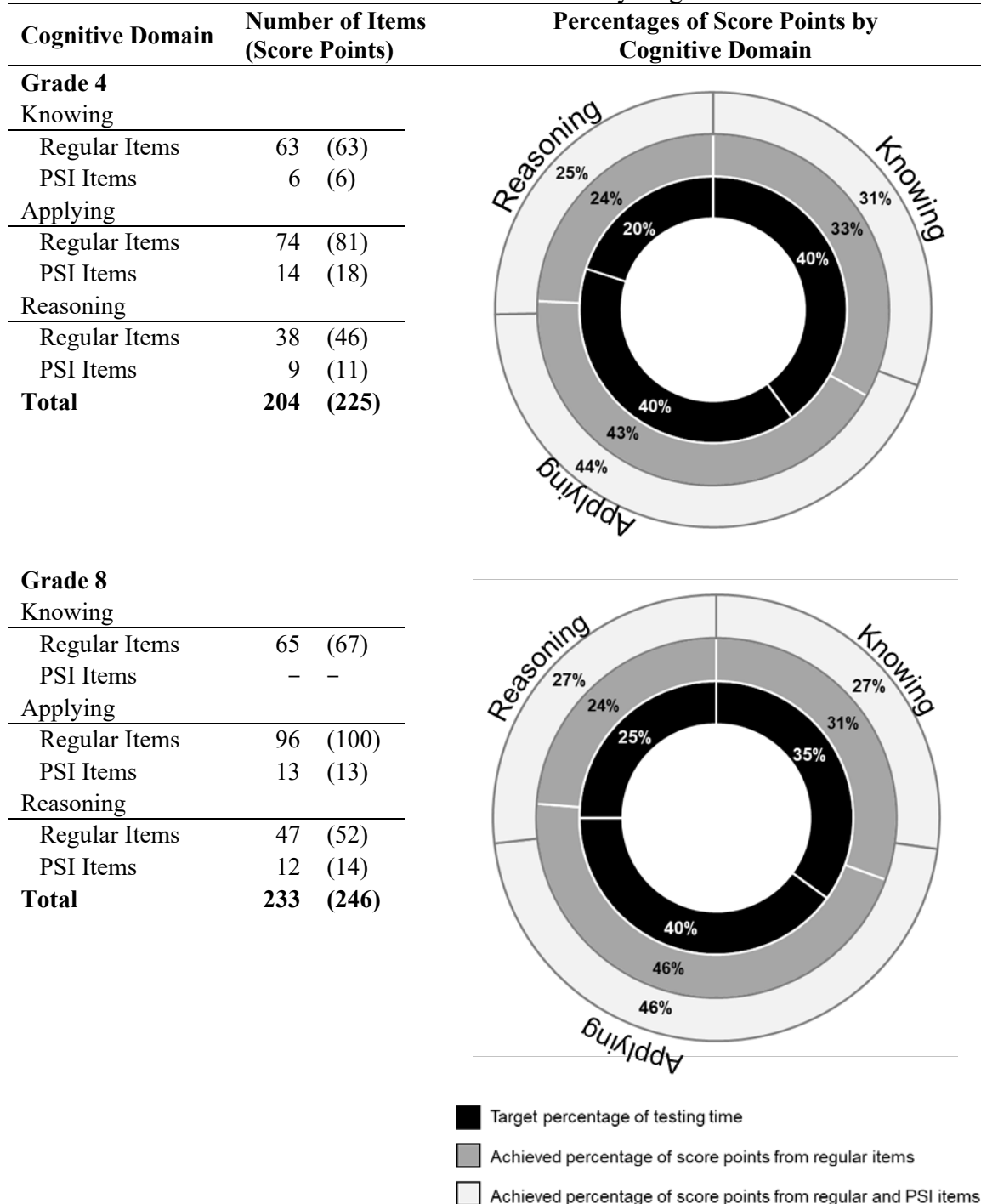
Because percentages are rounded to the nearest whole number, some results may appear inconsistent.

At the fourth grade, adding the PSI items to the regular eTIMSS mathematics items resulted in a four percentage point increase in the percentage of score points in the number content domain, bringing the coverage of this domain closer to the framework specifications, as well as a three percentage point decrease in coverage of measurement and geometry and a one percentage point decrease in coverage of data. Although the addition of the fourth grade PSIs caused some variation in the percentage of score points in each content domain, the small magnitude of these changes indicates that the PSIs did not meaningfully alter the content domain coverage provided by the assessment. At the eighth grade, the percentage of assessment score points allocated to the content domains was the same with and without the PSIs.

Cognitive Domain Coverage

Exhibit 4.4 shows the number of items and score points in the fourth and eighth grade mathematics assessments by framework cognitive domain (knowing, applying, and reasoning) and item type (regular or PSI) as well as multi-level pie charts. The center rings of the pie charts show the target percentage of testing time allocated to each cognitive domain in the framework, the middle rings show the achieved percentage of score points for the regular eTIMSS mathematics items, and the outer rings show the achieved percentage of score points for the regular and PSI items together.

Exhibit 4.4: eTIMSS 2019 Mathematics Assessments by Cognitive Domain



The fourth grade eTIMSS 2019 mathematics assessment included 175 regular items, worth a total of 190 score points, and 29 PSI items, worth a total of 35 score points. The eighth grade eTIMSS 2019 mathematics assessment included 208 regular mathematics items, worth a total of 219 score points, and 25 PSI items, worth a total of 27 score points.

Because percentages are rounded to the nearest whole number, some results may appear inconsistent.

The achieved percentages of score points from regular items in the fourth and eighth grade eTIMSS 2019 mathematics assessments were similar to the target percentages of testing time for the cognitive domains specified in the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017), but were less consistent with the framework specifications than the achieved content domain coverage. At both grades, the regular items provided more coverage of the applying domain and less coverage of the knowing domain than was specified in the framework. Although these achieved percentages deviate from the target specifications, the cognitive domain coverage of the regular items was still reasonably consistent with the intentions in the framework.

At both grades, consistent with the goal of the PSIs to assess problem solving, the majority of the items and score points in the PSIs addressed skills in the applying and reasoning domains. At the fourth grade, 23 items and 29 score points, or approximately 80 percent of the score points in the mathematics PSIs, were from items measuring students' applying and reasoning skills. At the eighth grade, the PSIs were exclusively comprised of items in the applying and reasoning domains, with about half the score points in the tasks allocated to each of these two domains. The eighth grade PSIs included 13 items and 13 score points in the applying domain and 12 items and 14 score points in the reasoning domain. In the reasoning domain, this was twice the percentage of score points specified in the framework for the regular items.

However, because the PSIs are such a small percentage of the full eTIMSS 2019 mathematics assessments, including the PSIs in the eTIMSS 2019 mathematics assessments resulted in only a slight increase in the percentage of score points in the

applying and reasoning domains at both grades. The fourth grade PSIs increased coverage of applying and reasoning by one percentage point in each domain, and the eighth grade PSIs increased coverage of the reasoning domain by three percentage points.

Validity of Response Process

Overall, the sources of qualitative and quantitative data collected during and after the eTIMSS 2019 Field Test provided evidence that the field test was conducted as planned and elicited the intended responses processes from students. The following sections discuss the functionality and usability of the instruments, student engagement in the assessment, and the scoring reliability for the constructed response items within the mathematics PSIs.

Several items in the eTIMSS 2019 Field Test Survey Activities Questionnaire completed by NRCs were considered in this dissertation (see Appendix B). NRCs from 22 eTIMSS countries responded to the survey. The results of the eTIMSS Student Questionnaire items discussed herein are provided in Appendix C.

Functionality and Usability

In appraising the functionality and usability of the eTIMSS assessments, this dissertation considered the eTIMSS Player that delivered the assessment items and recorded the students' responses, tools (ruler and calculator), and item types, as well as the clarity and utility of the directions and test administrator script. These aspects of the assessment were evaluated based on: 1) students' reports in the eTIMSS Student Questionnaire about difficulties experienced when taking the test, 2) NRCs' responses to

the Field Test Survey Activities Questionnaire, 3) NRCs' feedback on the PSIs collected after the field test, and 4) the author's observations of several field test testing sessions.

Using a variety of sources of information to evaluate the functionality and usability of the eTIMSS instruments not only helped to learn more about the impact of the technology used in eTIMSS on students' interactions with the test content, but also prompted improvements to the mathematics PSIs, eTIMSS Systems, directions, and manuals for eTIMSS 2019 Data Collection. The improvements made following the field test are discussed to demonstrate TIMSS' next steps in promoting response process validity.

eTIMSS Player

The eTIMSS Player was generally reliable in delivering the eTIMSS 2019 Field Test to students, capturing students' responses, and enabling test administrators to upload the data to the IEA's servers as planned. In the eTIMSS Student Questionnaire, approximately 95 percent of students at both grades reported that they were able to complete the test without any computer or tablet issues that required re-starting the test.

Still, all of the NRCs who responded to the Survey Activities Questionnaire reported that some intermittent technical issues arose during the field test testing sessions. Six NRCs reported that the eTIMSS Player occasionally froze or crashed during testing sessions, but that these freezes/crashes were easily rectified by restarting the device or player, or navigating away from the screen and back again. In these cases, students were able to continue the assessment with minimal disruption and no data were lost. However, several NRCs also reported occasional technical issues with the eTIMSS Player that did

result in a loss of data. Five NRCs reported that the player sometimes skipped over the second half of the assessment, preventing students from completing either the mathematics or science items in their assigned block combination. One NRC reported that in several instances the images within the items did not load, which prevented some students from taking the test.

NRCs also reported several issues related to the interaction between the eTIMSS Player and the devices used for field test data collection. Three methods were offered for administering the field test—1) individual PCs with the eTIMSS Player on USB sticks, 2) individual tablets with the eTIMSS Player software installed, and 3) using a central PC or Chromebook as a local server that delivers the eTIMSS Player to students’ PCs/Chromebooks via the school’s Local Area Network (LAN). Most countries used either the individual PC method, the individual tablet method, or a mix of the two. In the Survey Activities Questionnaire, nine NRCs from countries using the individual PC approach on school-owned computers reported experiencing issues with anti-virus software, available memory space, or software updates occurring during testing sessions. Two NRCs using individual tablets and two NRCs using the server method reported that the player froze more often when using these methods than when they used the individual PC method.

Following the field test, staff at IEA Hamburg investigated the reported issues and continued to work on increasing the stability of the eTIMSS Player across all three administration methods. Administering the field test also helped NRCs become familiar

with the eTIMSS Player and select the most appropriate administration method for their country for main data collection, which should help reduce the frequency of these issues.

eTIMSS Ruler Tool

Across both grades, there were a total of five mathematics items using the eTIMSS ruler tool. For all of these items, the ruler tool was activated and oriented horizontally on the screen when students arrived at the item. At the fourth grade, two of the ruler tool items required measuring a horizontal length and applying a scale to the measured length, one item required measuring a diagonal length (i.e., turning the ruler) and applying a scale, and one item required measuring a vertical length. At the eighth grade, there was one item within a PSI involving the ruler tool, which also required measuring a horizontal length and applying a scale.

Exhibit 4.5 shows the international average percent correct statistics for the fourth and eighth grade mathematics items using the ruler tool in the eTIMSS 2019 Field Test. For the four items that required applying a scale to correctly answer the item, a diagnostic score code was used to track the number of students who measured correctly, but did not apply the scale. The column headed “percent measured correct” in Exhibit 4.5 includes all students who provided a correct measurement for the length.

Exhibit 4.5: International Average Item Statistics for Mathematics Ruler Tool Items in the eTIMSS 2019 Field Test

Item Description	Format	Percent Correct	Percent Measured Correct	Percent Omitted
Grade 4				
Horizontal length and scale (1 cm = 4 m)	CR	28.9	83.0	1.0
Vertical length	MC	80.5	–	0.6
Horizontal length and scale (1 cm = 20 m)	CR	28.0	74.0	2.8
Diagonal length and scale (1 cm = 20 m)	CR	24.2	62.5	3.1
Grade 8				
Horizontal length and scale (1 cm = 100 cm)	CR	35.1	54.7	3.2

CR = constructed response; MC = multiple-choice. Two of the fourth grade items— Horizontal length and scale (1 cm = 4 m) and Vertical length—and the one eighth grade item were selected for eTIMSS 2019 Data Collection.

At both grades, the international average percent correct statistics provided evidence that students were successful in measuring with the ruler tool. Across the four fourth grade items in the field test, more than 60 percent of the students provided a correct measurement for the length to be measured and only a small number of students did not respond. At the eighth grade, more than half the students provided a correct measurement on the one ruler tool item in the field test. The eighth grade ruler tool item was situated in a more complex context within a PSI, which may have influenced the item difficulty.

Despite the satisfactory performance at the fourth grade, six NRCs reported in the Survey Activities Questionnaire that some fourth grade students experienced difficulties operating the ruler tool, particularly in items that required turning the ruler from the horizontal position to measure diagonal and vertical lengths. During the testing session observations, the author also observed several students repeatedly clicking on the ruler before understanding how to click and drag to move or turn the ruler. Based on these reports, TIMSS added instructions for the ruler tool to the general directions for main

data collection so that students who encounter the small number of items involving the ruler will not waste time or become frustrated trying to use it.

eTIMSS Calculator

At the eighth grade, students were provided with an on-screen calculator tool that they could access for any item by clicking the calculator icon at the bottom of the screen. The calculator tool functioned as a standard four-function calculator and also included a square root button. There were no reported issues with the functionality of the calculator, with the exception of two NRCs' reports in the Survey Activities Questionnaire that the tool did not work for a small number of students during testing sessions. The author did not see any issues with the calculator tool in the observed testing sessions and witnessed students who chose to use the tool fluently alternating between the calculator and scratch paper while solving problems.

Two NRCs reported in the Survey Activities Questionnaire that some students were confused by the functionality of the tool because it was not thoroughly explained in the directions and in many cases differed from the more complex calculators students in their countries were accustomed to using in class (e.g., scientific or graphing calculators). Based on this feedback, TIMSS added more details about how the calculator handles the order of operations to the eighth grade version of the eTIMSS directions for main data collection, including an example calculation for students to try before beginning the test.

Students' Interactions with eTIMSS Item Types

The eTIMSS mathematics field test instruments were comprised of a variety of item types, including traditional multiple-choice and constructed response, a wide assortment of enhanced item types (drop-down menus, selection, drag and drop, and

sorting) and two enhanced item formats specially designed for items within the mathematics PSIs—a line drawing tool and “sliders” used in several items involving positioning points on a line.

Exhibit 4.6 shows the number of regular and PSI items in the fourth and eighth grade mathematics field test instruments by item type. Students’ interactions with all of the field test items were considered in evaluating the response process validity of the eTIMSS item formats, although only about half were selected for main data collection.

Exhibit 4.6: Number of Regular and PSI Mathematics Items in the eTIMSS 2019 Field Test by Item Type

Item Type	Number of Items in the eTIMSS 2019 Field Test					
	Grade 4			Grade 8		
	Regular	PSI	Total	Regular	PSI	Total
Multiple-Choice	55	3	58	55	1	56
Number Pad	42	19	61	57	24	81
Keyboard	6	4	10	24	11	35
Drop-down Menu	3	—	3	3	4	7
Selection	9	3	12	7	—	7
Drag and Drop	12	7	19	12	—	12
Sorting	—	2	2	—	—	—
Line Drawing*	—	7	7	—	1	1
Sliders*	—	2	2	—	2	2
Total	127	47	174	158	43	201

*Item type was only available for items within the PSIs.

Counts reflect the total number of items in the eTIMSS 2019 Field Test instruments. Approximately half of these items were selected for eTIMSS 2019 Data Collection.

At the fourth grade, the PSIs included a greater proportion of enhanced item types than the regular items. Across the fourth grade PSIs, 45 percent of the items were enhanced, compared to 24 percent of the regular mathematics items. At the eighth grade, there was a smaller percentage of enhanced item types across both the regular and PSI items (14% and 16%, respectively). However, the eighth grade PSIs included more

interactive features (e.g., a video, an interactive graph, an interactive table) that are not reflected in these counts.

Comparison of eTIMSS and paperTIMSS Field Test Item Statistics

The digital and paper versions of the regular mathematics items were designed to be equivalent with the exception of the response mode, so comparing the average percent correct for the eTIMSS and paperTIMSS versions of these items helped to detect potential disadvantages or advantages of the digital format.

Exhibit 4.7 presents the international average item statistics for the regular mathematics items in the eTIMSS 2019 Field Test by eTIMSS item type and mode of administration. Because different groups of countries responded to the digital and paper formats of the items in the field test, the data analysis team at the TIMSS & PIRLS International Study Center used the item percent correct statistics from TIMSS 2015 to estimate the “selection bias,” or difference in performance between the groups of countries choosing to participate in eTIMSS versus paperTIMSS in 2019. The column headed “adjusted difference in percent correct” shows the difference in eTIMSS and paperTIMSS countries’ average percent correct with this adjustment for selection bias. Negative differences indicate a mode effect favoring the paper format. To provide a fair comparison across modes, the PSI items and ruler tool items that were only included in the eTIMSS field test were excluded from this analysis.

Exhibit 4.7: International Average Item Statistics from the Regular Mathematics Items in the eTIMSS/paperTIMSS 2019 Field Test by eTIMSS Item Type and Mode of Administration

eTIMSS Item Type	Number of Items	eTIMSS International Average		paperTIMSS International Average		Adjusted Difference in Percent Correct*
		DIFF	DISC	DIFF	DISC	
Grade 4						
Multiple-Choice Item Types						
Traditional	55	52.1 (19.5)	0.40 (0.1)	44.6 (18.6)	0.38 (0.1)	0.2
Drop-down Menu	3	52.7 (8.4)	0.36 (0.1)	57.4 (1.8)	0.45 (0.1)	−12.0
Constructed Response Item Types						
Number Pad	42	44.8 (20.4)	0.46 (0.1)	36.7 (18.0)	0.46 (0.1)	−2.1
Keyboard	6	23.3 (16.7)	0.40 (0.1)	20.0 (17.6)	0.36 (0.1)	−6.9
Selection	9	39.1 (22.6)	0.38 (0.0)	33.6 (23.6)	0.41 (0.0)	−4.7
Drag and Drop	12	68.3 (22.0)	0.33 (0.1)	55.6 (20.9)	0.46 (0.1)	2.6
Overall	127	49.0 (21.7)	0.41 (0.1)	41.5 (20.1)	0.41 (0.1)	−1.5
Grade 8						
Multiple-Choice Item Types						
Traditional	55	45.8 (13.8)	0.40 (0.1)	36.1 (12.3)	0.34 (0.1)	−6.0
Drop-down Menu	3	37.8 (9.0)	0.33 (0.2)	20.2 (12.5)	0.41 (0.2)	1.9
Constructed Response Item Types						
Number Pad	57	33.0 (14.7)	0.50 (0.1)	20.9 (11.7)	0.43 (0.1)	−8.7
Keyboard	24	22.4 (14.7)	0.47 (0.1)	14.6 (11.2)	0.39 (0.1)	−13.0
Selection	7	34.7 (19.3)	0.37 (0.1)	21.8 (11.8)	0.35 (0.1)	−7.9
Drag and Drop	12	35.3 (17.2)	0.36 (0.1)	26.9 (9.7)	0.45 (0.1)	−12.4
Overall	158	36.2 (16.6)	0.44 (0.1)	25.7 (14.1)	0.39 (0.1)	−7.8

*Difference between international average percent correct across eTIMSS and paperTIMSS with an adjustment based on country selection bias, estimated based on the 20 fourth grade countries and 15 eighth grade countries that participated in TIMSS 2015. The adjustment was 7.2 percentage points for multiple-choice item types and 10.2 percentage points for constructed response item types at the fourth grade and 15.8 percentage points for multiple-choice item types and 20.7 percentage points for constructed response item types at the eighth grade. Negative differences indicate a mode effect favoring the paper format.

() Standard deviations appear in parentheses. Because of rounding, some results may appear inconsistent.

At the fourth grade, eTIMSS countries' adjusted average percent correct on the traditional multiple-choice items, number pad items, selection items, and drag and drop items were all within 5 percentage points of the paperTIMSS countries' performance on the paper-based versions of these item types. These results suggest that the digital format of these eTIMSS item types did not substantially impact fourth grade students'

interactions with the test content. For the very few items using the keyboard and drop-down menus, the differences in the adjusted percent correct across eTIMSS and paperTIMSS indicate that these formats were somewhat more challenging in eTIMSS.

Although some disadvantage was expected for the six eTIMSS items involving typing based on the results of the eTIMSS Item Equivalence Study (Fishbein et al., 2018), the difference for the three items involving drop-down menus was surprising, because this format is a variant of traditional multiple-choice. However, further investigation of how the drop-down menus were used in these items revealed that the layouts were problematic, rather than the drop-down menus themselves. Following the field test, these items were revised to address the layout issues.

At the eighth grade, the adjusted differences in percent correct indicated that all item types, with the exception of the three drop-down menus, were more challenging in eTIMSS. This data suggests a larger mode effect between eTIMSS and paperTIMSS at the eighth grade than the fourth grade.

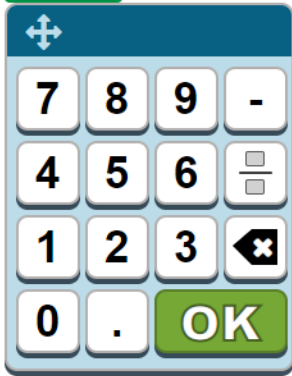
Qualitative Analysis of Students' Interactions with the eTIMSS Item Types

Students' responses to the eTIMSS Student Questionnaire, NRCs reports via the Survey Activities Questionnaire and feedback on PSIs, and the author of this dissertation's observations of testing sessions were used to further investigate students' interactions with the constructed response item types in the eTIMSS 2019 Field Test. This included the number pad and keyboard, selection, drag and drop, and sorting, as well as the line tool and slider features used exclusively for PSI items. The following sections present the result for each constructed response item type.

Numeric Constructed Response Items

For all constructed response items requiring a numeric answer (i.e., integers, decimals, and fractions), the eTIMSS students responded using the eTIMSS number pad. Exhibit 4.8 demonstrates the appearance and functionality of the eTIMSS number pad, which included the digits 0 to 9, a decimal point, a negative sign, a fraction button, and enter and backspace buttons. At both grades, this was the most common item format in the mathematics field test instruments (approximately 30% of the fourth grade items and 40% of the eighth grade items).

Exhibit 4.8: Functionality of the eTIMSS Number Pad for Numeric Constructed Response Items

<i>Initial appearance of response space</i>	<i>Number pad becomes activated when students click on the response space</i>	<i>Students can type positive and negative numbers, fractions, and decimals</i>
Answer: <input type="text"/>	Answer: <input type="text"/> 	Answer: <input type="text" value="1"/> Answer: <input type="text" value="-1"/> Answer: <input type="text" value="1/2"/> Answer: <input type="text" value="0.1"/>

Students' responses to the eTIMSS Student Questionnaire indicated that the majority of students at both grades did not have trouble using the number pad. At the fourth grade, 90 percent of students reported that they did not have trouble with the number pad and 79 percent of the eighth grade students reported that they did not have trouble. The NRCs did not report any issues with the number pad at either grade.

During the testing session observations, the author noted that the number pad was immediately intuitive for eighth grade students, but many fourth grade students required assistance from the test administrator when practicing the number pad during the eTIMSS directions. However, all fourth grade students appeared to become increasingly dexterous in using the tool as they progressed through the assessment. After completing several items with the number pad, all fourth grade students were seen entering numbers quickly and instinctively moving the number pad around the screen to see any parts of the item it obscured.

Typed Constructed Response Items

For the few mathematics constructed response items requiring a written response, students typed their answers in a designated response space using on the standard keyboard for the device on which they were taking eTIMSS. At both grades, 82 percent of students reported via the eTIMSS Student Questionnaire that typing their answers was not difficult and no issues were reported about the functionality of this item type. In the feedback collected for the PSIs, one NRC suggested that the keyboard icon next to the on-screen typing area be removed because students did not have to click it to begin typing. Based on this comment, this icon was removed for main data collection.

In the Survey Activities Questionnaire, one NRC reported that some fourth grade students were slowed down by having to type their answers. During the testing session observations, the author also noted that some fourth grade students typed very slowly, but none appeared to be frustrated by having to type. At the eighth grade, students were more proficient in typing, but encountered some difficulties typing responses involving

mathematical symbols. In the Survey Activities Questionnaire, one NRC reported that students were confused about how to express mathematical symbols that do not appear on a standard keyboard (e.g., multiplication, division, and exponents) and suggested adding standard conventions for these symbols to the eTIMSS directions. As countries were scoring student responses to the field test, six additional NRCs contacted the TIMSS & PIRLS International Study Center with questions about how to score responses with unconventional mathematics notation (e.g., x2 instead of x^2). To address this issue, TIMSS added instructions for typing mathematical symbols to the eTIMSS directions for main data collection.

Selection Items

No issues were reported with the majority of the selection items used in the field test in which students could choose an unlimited number of options to respond to the question (e.g., click **all** the shapes). However, in the feedback collected for the PSI items, two NRCs reported that the three mathematics selection items that were designed with a limited number of options to be selected (e.g., click **two** shapes) were confusing for students, especially at the fourth grade. For these selection items, clicking an additional part above the designated limit resulted in the first part that was chosen to automatically become unselected, making it more challenging to convey a final answer once the maximum number of selectable options was exceeded. Following the field test, TIMSS changed all mathematics selection items to the unlimited format to mitigate such issues.

Drag and Drop Items

The drag and drop items in the field test varied widely in the elements to be dragged (e.g., labels, images) and the appearance of the “drop zone” where the draggable elements could be placed (e.g., a table, a pictograph, or a blank space in a number sentence). Still, with a standard procedure for dragging, the diversity of the drag and drop items did not present any notable issues for students. In the eTIMSS Student Questionnaire, approximately 80 percent of students at both grades reported that they did not have trouble dragging objects and the NRCs did not report any issues with the standard drag and drop items. During the testing session observations, the author noted that larger draggable elements appeared to be easier for students to move, especially with the stylus. However, all students at both grades seen responding to drag and drop items appeared able to construct their answers with minimal difficulty.

One series of drag and drop items within a fourth grade mathematics PSI that required not only dragging, but also rotating draggable tiles to make a design, was found to be problematic in the field test. To position a tile, students needed to drag it into place, click it again to get a turning arrow to appear on the corner of the tile, then click, hold, and drag the turning arrow to turn the tile. In the feedback collected on the PSIs, seven NRCs reported that many students had trouble turning the tiles. Based on this feedback, no drag and drop items with rotation were brought forward to data collection.

Sorting Items

The sorting format was only used for two fourth grade mathematics items in the field test. This component functioned similarly to drag and drop, but as students dragged any of the elements to be sorted, the other sortable elements shifted to account for the

move. The NRCs did not report any issues with the two mathematics items using this format for the field test, but the two sorting items had exceptionally high omit rates (an average of 27.4% across the two items), suggesting that some students may have experienced difficulties with this component or were unsure of how to use it. For data collection, TIMSS added more instructions on these two screens to help direct students to the response spaces. However, in general, sortable items have been discontinued.

Slider Items

Sliders were used for two PSI items at each grade. At the fourth grade, this item type was used to ask students to show values on number lines by dragging pointers along the lines. At the eighth grade, it was used to ask students to position data points on graphs by sliding the points up and down the y-axes. In the feedback collected for the PSIs, one NRC reported that fourth grade students did not recognize the sliders as interactive elements. Again, this issue was clearly reflected in the percentage of students that omitted these items in the field test. Across the two fourth grade slider items, the international average omit rate was 30 percent. Following the field test, TIMSS revised the fourth grade slider items to draw more attention to these unique response spaces and provided more explicit explanation of how to interact with them. At the eighth grade, the international average percentage of students that omitted the slider tool items was more consistent with the omit rate for the other item types within the PSIs (13.5%), indicating that most eighth grade students recognized and interacted with the tool.

Line Drawing Items

Two PSIs in the field test included items asking students to use a tool with “snap to grid” behavior to draw straight lines with their mouse, stylus, or finger on a square grid. Two NRCs reported that fourth grade students had trouble drawing on the grid with a mouse and one NRC reported that students were confused about the difference between the “undo” and “reset” buttons that were provided for student to either erase the last line they drew (undo) or clear the entire grid (reset). During the testing session observations, the author noted that students using a stylus at both grades engaged in some trial and error in determining how to use the tool, but became comfortable using it after drawing several lines. Beyond students’ issues, technical difficulties prevented students’ responses to the line drawing items in the field test from being reliably displayed in the IEA’s Online Scoring System. However, based on the lessons learned in developing and field testing this item type in the PSIs, staff at the TIMSS & PIRLS International Study Center designed an improved line drawing tool that could be used for regular eTIMSS items in main data collection.

Developing Concise Directions

At the beginning of the eTIMSS field test testing sessions, test administrators followed a script to lead students through detailed directions explaining how to navigate the assessment, respond to the traditional and enhanced item types, and use the eTIMSS tools. The eTIMSS directions for the field test included a series of example items for students to practice using each item type and test administrators were asked to ensure that all students successfully completed the example items before beginning the assessment. Because of the wide range in the level of computer skills across the target population, the

eTIMSS directions and test administrator script used in the field test were designed to be thorough enough to safeguard against unintended bias associated with students' lack of computer skills.

Students' performance across the item types in the eTIMSS 2019 Field Test (shown Exhibit 4.7) indicated that the directions were generally successful in ensuring that students were able to respond to the wide variety of item formats included in the assessment. However, feedback from NRCs and the author's observations of testing sessions indicated that the directions were overly detailed for most students, particularly at the eighth grade. In the Survey Activities Questionnaire, three NRCs reported that the field test directions took too long to complete (20 to 30 minutes), causing some students to lose focus before the test began. Three more NRCs echoed this concern in the feedback collected for the PSIs, and the author of this dissertation saw students becoming very restless during the directions in the observed testing sessions. Based on this feedback, TIMSS substantially reduced the amount of text in the eTIMSS directions and test administrator script after the field test and encouraged NRCs to adapt the script to be suitable for students in their countries if needed.

At the same time, six NRCs requested via the Survey Activities Questionnaire or their feedback on the PSIs that TIMSS add more details to the eTIMSS directions to more thoroughly explain the ruler and calculator tools, how to type mathematical symbols, and how to use the enhanced item types specific to the PSIs. For data collection, TIMSS added instructions to the eTIMSS directions for using the ruler tool, more detail about the functionality of the eTIMSS calculator (e.g., that it does not apply the order of operations

to multi-step calculations), and a brief note explaining how to type mathematical symbols on a keyboard (e.g., for multiplication, use the letter “x”). The NRCs reviewed the updated eTIMSS directions before the instruments were finalized for data collection to ensure that the most pressing issues had been addressed.

Students Found eTIMSS Engaging

The eTIMSS assessments and particularly the PSIs were designed to be more engaging and motivating for students than paper-based assessment. In the eTIMSS Student Questionnaire, the majority of students at both grades reported that they liked taking the assessment on a computer or tablet. At the fourth grade, 67 percent of students reported that they “liked it a lot” and 27 percent of students reported that they “liked it a little,” resulting in a total of 94 percent of students with positive attitudes toward taking the test on a digital device. At the eighth grade, 82 percent of students expressed positive attitudes toward taking the test on a computer/tablet, with 44 percent of students reporting that they “liked it a lot” and 38 percent reporting that they “liked it a little.” In the feedback collected from NRCs following the field test, four NRCs reported that their students found the PSIs to be challenging, but fun and engaging.

During the observed testing sessions, the author noted that almost all students at both grades appeared to be highly engaged in the assessment and giving their full effort on all items. At the fourth grade, a total of three students across the observed testing sessions were seen playing with the interactive features and not answering the questions, and no students were seen going off task in the eighth grade classes. The author also observed that students taking the PSIs appeared to spend more time on each screen in the

assessment and be less willing to move on to the next question until they were satisfied with their answer than students taking the regular eTIMSS mathematics items. This difference may be related to the amount of information on the PSI screens, the item difficulty, or because students were engaged in and motivated by the problem contexts.

Further, the author noted that the students who were able to successfully answer most items in the PSIs became increasingly invested in the problem scenarios as they progressed through them. However, students who appeared to be struggling with the mathematics content in the PSIs appeared to become fatigued or less engaged toward the end of the testing sessions, suggesting that the extended contexts could be difficult for lower achieving students.

In the feedback collected on the mathematics PSIs, the NRCs' most common critique was that the tasks required too much reading, which caused some students to become discouraged or give up. Based on this feedback, staff at the TIMSS & PIRLS International Study Center and expert consultants scrutinized the text in each PSI selected for main data collection to reduce the reading load as much as possible.

Time on Screen Varied by Cognitive Domain

The timing data collected during the eTIMSS 2019 Field Test was used to further investigate student engagement and the agreement between the hypothesized cognitive demands of the items and students' actual time on task.

Exhibit 4.9 shows the international average number of seconds students spent on each screen with mathematics items in the eTIMSS 2019 Field Test by the cognitive domain and total score points on the screen. Results are shown for the regular eTIMSS

screens compared to the PSI screens. For the PSI screens, the average time per screen is shown separately for the two positions in which the task appeared under the assessment design used in the field test—as the first block or second block in the mathematics part of a block combination. Exhibit 4.10 shows the results for the eighth grade.

Exhibit 4.9: Average Time per Screen in the eTIMSS 2019 Field Test by Cognitive Domain and Score Points – Grade 4

	Regular Blocks		Number of Screens	PSI Blocks		Average Time per Screen in Seconds
	Number of Screens	Average Time per Screen in Seconds		Average Time per Screen in Seconds		
				Position 1	Position 2	
Knowing Items						
1 point	31	45.9 (16.8)	1	58.4 –	35.4 –	45.9 (16.5)
2 points	3	82.5 (24.4)	1	88.0 –	78.2 –	82.6 (19.9)
All items	34	49.1 (20.1)	2	73.2 (20.9)	62.9 (30.3)	50.0 (20.3)
Applying Items						
1 point	52	63.3 (19.4)	3	74.1 (31.7)	41.3 (8.6)	63.0 (19.3)
2 points	7	96.2 (37.3)	3	126.9 (45.9)	69.6 (25.3)	96.8 (34.6)
3 points	–	– –	7	133.9 (46.1)	69.3 (23.6)	101.6 (34.1)
All items	59	67.2 (24.3)	13	118.5 (47.2)	62.9 (24.3)	71.4 (27.7)
Reasoning Items						
1 point	14	81.8 (30.7)	6	103.8 (33.9)	61.0 (37.1)	82.0 (31.0)
2 points	9	125.1 (49.5)	5	127.0 (64.0)	65.4 (43.3)	114.8 (50.0)
3 points	–	– –	3	166.9 (49.6)	71.3 (31.1)	119.1 (40.2)
All items	23	98.8 (32.8)	14	125.6 (51.9)	64.8 (35.7)	97.4 (42.3)

Average time is the average of 30 participating countries' average time per screen. Timing data for the United States were not collected in the field test.

For PSI items, the average time per screen is provided separately for the two positions in which it appeared under the assessment design used in the field test—as the first block or second block in the mathematics part of a block combination.

() Standard deviations appear in parentheses.

Exhibit 4.10: Average Time per Screen in the eTIMSS 2019 Field Test by Cognitive Domain and Score Points – Grade 8

	Regular Blocks		Number of Screens	PSI Blocks		Average Time per Screen in Seconds
	Number of Screens	Average Time per Screen in Seconds		Average Time per Screen in Seconds		
				Position 1	Position 2	
Knowing Items						
1 point	38	57.4 (19.3)	–	–	–	57.4 (19.3)
2 points	4	79.3 (22.0)	1	89.6	75.9	80.0 (19.1)
All items	42	59.5 (20.3)	1	89.6	75.9	60.0 (20.4)
Applying Items						
1 point	63	64.4 (23.7)	5	100.5 (27.1)	70.2 (20.4)	65.9 (24.2)
2 points	6	98.7 (16.3)	1	81.1	61.4	94.7 (18.2)
3 points	–	–	1	172.6	111.1	142.0
All items	69	67.4 (25.0)	7	108.0 (36.8)	74.8 (23.4)	69.6 (26.3)
Reasoning Items						
1 point	18	84.5 (27.1)	1	11.1	7.4	80.6 (31.5)
2 points	13	111.6 (53.2)	8	192.2 (51.4)	128.7 (43.1)	130.2 (54.4)
3 points	1	184.7	5	166.1 (47.5)	109.7 (29.8)	145.7 (39.0)
4 points	–	–	1	273.6	137.3	205.5
All items	32	98.7 (43.9)	15	176.9 (69.1)	114.9 (46.5)	113.7 (40.7)

Average time is the average of 20 participating countries' average time per screen. Timing data for the United States were not collected.

For PSI items, the average time per screen is provided separately for the two positions in which it appeared under the assessment design used in the field test—as the first block or second block in the mathematics part of a block combination.

() Standard deviations appear in parentheses.

At both grades, for both regular and PSI screens, the average amount of time per screen increased with the level of cognitive demand and number of score points per screen. At the fourth grade, students spent an average of approximately 20 seconds more per screen with each increase in level of cognitive demand (50.0 seconds on knowing screens, 71.4 seconds on applying screens, and 97.4 seconds on reasoning screens). At the eighth grade, students spent an average of nine seconds more on applying screens than knowing screens (69.6 seconds on applying screens versus 60.0 seconds on knowing screens) and an average of approximately 40 seconds more on reasoning screens than

applying screens (113.7 seconds on reasoning screens). These results show that the cognitive classifications and point values of the field test items were consistent with students' interactions with the items, providing validity evidence for these classifications and evidence that students were engaged in the assessment.

At both grades, on average, students spent more time on all PSI screens when they appeared first in the mathematics part of a block combination than when they appeared second. This pattern held across all cognitive domains and numbers of score points, but the difference in the average time on screen between the first and second positions increased with the cognitive complexity of the item. At the fourth grade, students spent approximately twice the amount of time on applying and reasoning screens within the PSIs when they appeared first compared to when they appeared second. At the eighth grade, students spent an average of about 30 seconds more on applying screens and 60 seconds more on reasoning screens when they appeared first compared to when they appeared second. These differences in average time on task across the two positions show a position effect for the PSIs that may have resulted from extra time to become familiar with the PSI assessment format or the difficulty of the items.

At the same time, these results indicate that the PSIs were engaging and motivating, as they compelled students to spend time working through challenging problems. Using the average time per screen from the first position for comparison, students at both grades spent considerably more time on PSI screens than on regular mathematics screens across all cognitive domains. At the fourth grade, students spent an average of 24 seconds more on knowing screens, 51 seconds more on applying screens,

and 26 seconds more on reasoning screens within the PSIs compared to the regular screens. At the eighth grade, student spent an average of 41 seconds more on applying screens and over a minute (81 seconds) more on reasoning screens compared to the regular eighth grade mathematics screens.

Together with the NRCs' reports and the author's observations, these results indicate that the PSIs provided the necessary scaffolding to support students in persevering through challenging mathematics items.

PSI Items were Scored Reliably

At both grades, the majority of the items in the mathematics PSIs in the eTIMSS 2019 Field Test were constructed response items. TIMSS' efforts to ensure reliable scoring of both the machine- and human-scored constructed response items in the PSIs for the eTIMSS 2019 Field Test proved to be very successful.

Exhibit 4.11 shows the number of machine- and human-scored constructed response items in the mathematics PSIs field tested at each grade. For the human-scored items, the international average percent of agreement across scorers was 97 percent for the 100 student responses to each item that were double-blind scored in each country.

Exhibit 4.11: Machine- and Human-Scored Constructed Response Items in Mathematics PSIs in the eTIMSS 2019 Field Test

Grade	Number of Construct Response Items			International Average Percentage Agreement on Scores for Human-Scored Items
	Total	Machine- Scored	Human- Scored	
Grade 4	37	33	5	97
Grade 8	37	25	12	97

eTIMSS line drawing items were excluded from these counts.

Approximately two-thirds of these constructed response items were machine-scored, including all items using the number pad and eTIMSS enhanced item types. After

several iterations of review and verification among staff at the TIMSS & PIRLS International Study Center and IEA Hamburg, the scoring specifications and resulting scores assigned to all student responses to the machine-scored constructed response items within the field test PSIs were verified. The verification process provided evidence that the machine-scored constructed response item formats in the mathematics PSIs could be scored reliably and helped to inform next steps in refining data capture and scoring specifications for main data collection.

The items involving typed responses were sent to the IEA's Online Scoring System and scored by the participating countries. They were successfully displayed for scorers and at both grades the responses to these items were scored with a high degree of reliability across all countries (shown in Exhibit 4.11), consistent with the scoring reliability for the regular eTIMSS mathematics items (98% at the fourth grade and 96% at the eighth grade).

Validity of Internal Structure

The data collected in the eTIMSS 2019 Field Test were used to conduct a preliminary evaluation of the internal structure of the eTIMSS 2019 mathematics assessments. First, the timing data collected during the field test were used to investigate potential speededness or position effects that could have impacted the measurement properties of the items. Next, using the field test item statistics for the selection of items that moved forward to eTIMSS 2019 Data Collection, the measurement properties of the regular and PSI mathematics items were compared to investigate potential differences in the psychometric properties of the PSI items, compared to the regular items. Then, the

correlation between countries' average percent correct on the regular items and PSI items was examined to begin to evaluate whether the two item types measured the same construct. Finally, a series of confirmatory factor analysis and item response theory models were used to investigate the underlying relationships among the regular and PSI items and students' mathematics ability to determine whether both item types could be validly reported together on a unidimensional scale.

Speededness and Position Effects

Exhibit 4.12 shows the average time students spent completing each mathematics item block in the eTIMSS 2019 Field Test with the number of screens and score points in the block. For the PSI blocks, the average time is shown separately for the two positions in which it appeared.

Exhibit 4.12: International Average Time per Block in the eTIMSS 2019 Field Test

Block	Grade 4		Grade 8			
	Number of Screens (Score Points)	Average Time in Minutes	Number of Screens (Score Points)	Average Time in Minutes		
Regular						
ME01	11 (14)	14.4 (1.0)	15 (16)	20.1 (0.4)		
ME02	12 (13)	11.0 (0.2)	14 (17)	15.4 (0.5)		
ME03	11 (13)	13.2 (0.4)	14 (16)	17.1 (0.7)		
ME04	10 (12)	11.1 (0.4)	15 (18)	15.8 (0.5)		
ME05	12 (14)	14.3 (0.3)	13 (15)	20.6 (0.9)		
ME06	12 (12)	13.6 (0.6)	14 (17)	14.6 (0.5)		
ME07	12 (16)	14.8 (0.7)	14 (17)	17.0 (0.6)		
ME08	12 (14)	10.9 (0.4)	14 (16)	15.6 (0.4)		
ME09	11 (13)	14.1 (0.6)	16 (19)	20.0 (0.4)		
ME10	13 (14)	14.3 (0.4)	14 (18)	15.5 (0.4)		
Average	12 (14)	13.2 (1.6)	14 (17)	17.2 (2.3)		
PSI						
		<u>Position 1</u>	<u>Position 2</u>		<u>Position 1</u>	<u>Position 2</u>
MI01	9 (16)	18.2 (1.2)	8.7 (0.5)	10 (12)	21.0 (1.0)	12.9 (0.8)
MI02	14 (19)	18.3 (0.7)	11.5 (0.6)	10 (19)	18.2 (0.8)	14.1 (0.7)
MI03	15 (23)	20.8 (0.4)	10.4 (0.3)	9 (17)	19.1 (1.6)	11.8 (0.9)
Average	13 (19)	19.1 (1.5)	10.2 (1.4)	10 (16)	19.4 (1.4)	12.9 (1.2)

Average time is the average across the participating countries' average time per screen based on 30 countries at the fourth grade and 20 countries at the eighth grade. Timing data for the United States were not collected.

For number of screens, score points appear in parentheses. For average time in minutes, standard deviations appear in parentheses.

For PSI items, the average time per screen is provided separately for the two positions in which it appeared under the assessment design used in the field test—as the first block or second block in the mathematics part of a block combination.

The average time students spent on the regular mathematics blocks at both grades was relatively consistent across all blocks. At the fourth grade, the average across the regular blocks was 13.2 minutes and the range was approximately 4 minutes. At the eighth grade, the average time was 17.2 minutes and the range was approximately 5 minutes. At both grades, the average time for all of the regular blocks was less than the time allocated per block under the assessment design, indicating that the blocks were an

appropriate length for the testing time and that for most students the eTIMSS 2019 Field Test was not a speeded test.

Under the block combination design for the regular blocks, the odd-numbered blocks (ME01, ME03, ME05, ME07, ME09) appeared first in the mathematics part of a block combination, followed by an even-numbered block (ME02, ME04, ME06, ME08, ME10). At both grades, the average time per block for all odd-numbered blocks was greater than the time per block for the even-numbered blocks, suggesting that it may take students some time to get comfortable with the eTIMSS Player, format of the items, or the specific device used to administer the assessment. Fourth grade students spent an average of 14.2 minutes on odd-numbered blocks compared to an average of 12.2 minutes on even-numbered blocks and eighth grade students spent an average of 19.0 minutes on odd-numbered blocks and an average of 15.4 minutes on even-numbered blocks. Still, these results indicate that the position of the regular mathematics blocks within the block combinations did not substantively effect the amount of time students spent on the items within a block or cause speededness.

The difference in the average amount of time for the PSI blocks in the first and second positions indicates that there was a position effect for the fourth grade PSI items in the eTIMSS 2019 Field Test. The average time across the fourth grade mathematics PSI blocks when they appeared first was nearly twice the average time for when they appeared second (19.1 minutes versus 10.2 minutes) and the average time for all three of the PSI blocks when they appeared first exceeded the time allocated for a block by at least one minute. At the eighth grade, the difference in average time across the first and

second positions was closer to the difference between the odd- and even-numbered regular blocks (approximately 7 minutes) and none of the average times per block exceeded the time allocated to it.

To further investigate the implications of this position effect on the items, the average time for each screen within the PSI blocks was compared across the two positions in which it appeared. Double bar graphs of the average time by screen in the first and second position for each PSI block in the eTIMSS 2019 Field Test are provided in Appendix D. The average time per screen when it appeared first was consistently greater than the average time per screen when it appeared second, and the difference between the average times per screen generally became more pronounced towards the end of each block. This pattern suggests that the psychometric properties of the PSI items, and particularly those near the end of the blocks, were impacted by the assessment design.

The average percentage of students who did not reach all of the items in their assigned block combination (percent not reached) was also slightly higher in the PSI block combinations than the regular block combinations. At the fourth grade, the average percent not reached was 2 percentage points higher across the PSI block combinations than the regular block combinations (PSI not reached = 3.23%; Regular not reached = 1.25%). At the eighth grade, the average percent not reached was approximately 4 percentage points higher for the PSI block combinations (PSI not reached = 4.59%; Regular not reached = 0.91%). Although the percentage of students not reaching all items in a block was still relatively small for the PSI blocks, these percentages provide

additional evidence that students taking the PSIs may have been rushed to complete the test, whereas most students taking the regular mathematics items were not.

Measurement Properties of Items

Exhibit 4.13 presents the international average item difficulty (percent correct), discrimination (point-biserial correlation), and percent omitted for the fourth and eighth grade mathematics items in the eTIMSS 2019 Field Test that were selected for data collection. The results are reported by item format (multiple-choice and constructed response), as well as by item type (regular and PSI) to allow for reasonable comparisons across the two groups of items.

At both grades, the PSI items were more difficult than the regular mathematics items. The difference in the international average percent correct was approximately 9 percentage points at the fourth grade and 17 percentage points at the eighth grade, indicating that the PSI items were substantially more challenging, especially at the eighth grade. However, despite being more difficult, the PSI items had approximately the same international average item discrimination as the regular items. The international average item discrimination for both item types was very good—approximately 0.45 at the fourth grade and approximately 0.51 at the eighth grade—signifying that both the regular and PSI mathematics items were successful in differentiating between high and low achieving students.

Exhibit 4.13: International Average Item Statistics from the eTIMSS 2019 Field Test for Mathematics Items Selected for eTIMSS 2019 Data Collection

Item Format and Type	Number of Items	International Average Item Statistics		
		Average Item Difficulty	Average Item Discrimination	Average Percent Omitted
Grade 4				
Regular				
Multiple-choice	28	58.3 (15.5)	0.46 (0.1)	2.1 (1.3)
Constructed Response	48	46.1 (18.6)	0.45 (0.1)	5.3 (3.6)
Regular Overall	76	50.6 (18.4)	0.45 (0.1)	4.1 (3.3)
PSI				
Multiple-choice	2	41.0 (17.5)	0.37 (0.1)	13.2 (12.7)
Constructed Response	25	41.4 (18.3)	0.45 (0.1)	14.7 (11.2)
PSI Overall	27	41.4 (17.9)	0.45 (0.1)	14.6 (11.1)
Overall	103	48.2 (18.6)	0.45 (0.1)	6.9 (7.8)
Grade 8				
Regular				
Multiple-choice	24	50.2 (10.4)	0.50 (0.1)	2.4 (1.3)
Constructed Response	62	35.3 (15.0)	0.51 (0.1)	10.8 (8.2)
Regular Overall	86	39.5 (15.3)	0.51 (0.1)	8.5 (7.9)
PSI				
Multiple-choice	1	22.7 –	0.38 –	4.6 –
Constructed Response	21	22.8 (9.6)	0.51 (0.1)	16.3 (9.6)
PSI Overall	22	22.8 (9.4)	0.50 (0.1)	15.8 (9.7)
Overall	108	36.1 (15.8)	0.51 (0.1)	10.0 (8.8)

Multiple-choice includes traditional multiple-choice, compound multiple-choice, and drop-down menu items. Constructed response includes number pad, keyboard, selection, drag and drop, sorting, and slider items.

Only items in the eTIMSS 2019 Field Test that were selected for eTIMSS 2019 Data Collection are included in this analysis.

() Standard deviations appear in parentheses. Because of rounding, some results may appear inconsistent.

Internationally, on average, a higher percentage of students omitted PSI items than regular items. At the fourth grade, the average percent omitted across the PSI items was about 11 percentage points higher than the average percent omitted across the regular mathematics items. At the eighth grade, this difference was approximately 8 percentage points. Given the evidence of speededness and position effects detected in the analysis of the timing data, these higher omit rates across the PSI items may be in part attributed to

students running out of time and skipping over parts of items. However, it is difficult to parse out the cause of these omitted responses as the item difficulty, reading load, and technology also may have contributed to this difference between the regular and PSI items.

The PSI items were designed to address traditionally hard to measure areas of the framework and capitalize on technology beyond what was possible with the regular eTIMSS items, so some differences in the average item difficulty and percent omitted were anticipated. Although the psychometric properties of the PSI items in the eTIMSS 2019 Field Test did differ from the regular items in some respects, these results indicate that the PSI items were not vastly different from the regular mathematics items at either grade.

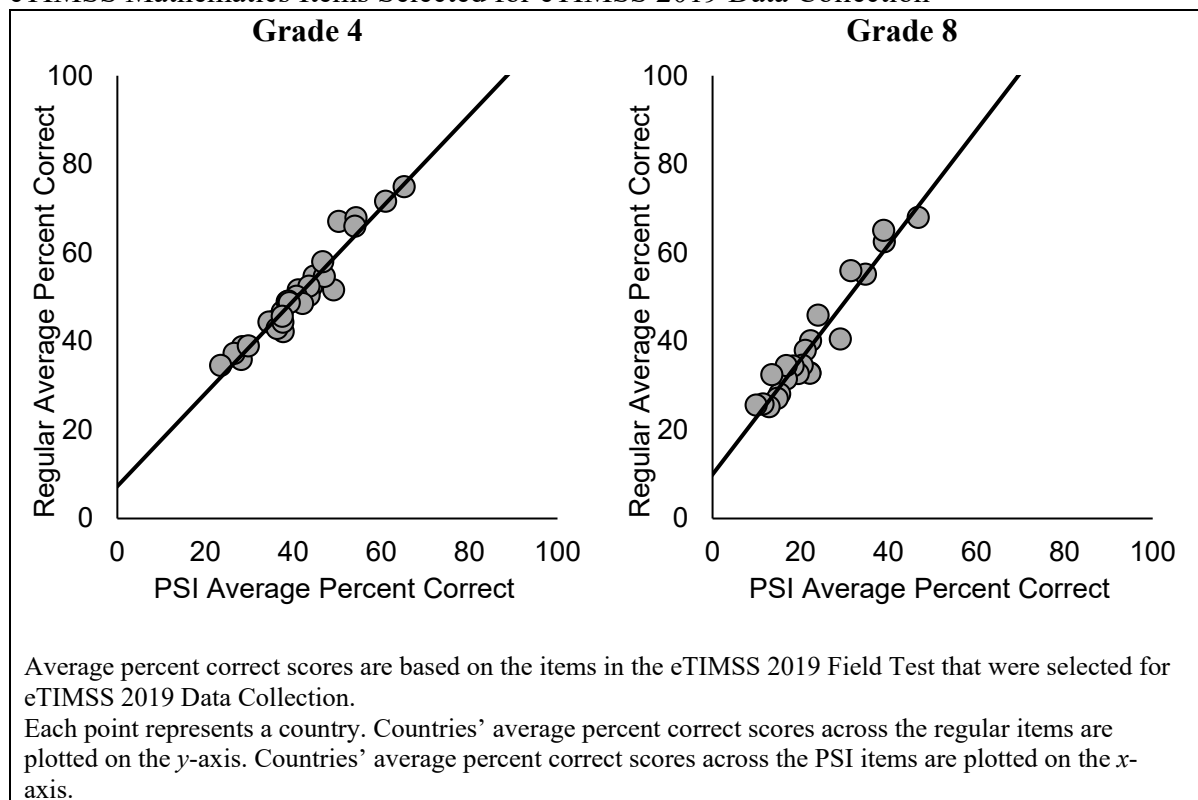
Following the field test, TIMSS simplified the most difficult items within each PSI task, added clearer explanations of enhanced item types and features, and further reduced the reading load in all of the PSIs selected for main data collection. These changes are expected to bring the PSI items even closer into line with the regular mathematics items by increasing the international average item difficulty, making the PSIs a more appropriate length for the testing time, and decreasing the percent of omitted responses.

Performance Consistency across Regular and PSI Items

Exhibit 4.14 shows countries' average percent correct across the regular items plotted against the average percent correct across the PSI items at the fourth and eighth grades. Consistent with the previous comparison of the international item statistics

(Exhibit 4.13), the average percent correct scores for each country were calculated based on the results from the eTIMSS 2019 Field Test for the items that were selected for main data collection. On the plots, each point represents a country, with the average percent correct on the regular items on the y -axis and the average percent correct on the PSI items on the x -axis. The linear line of best fit for the data is shown.

Exhibit 4.14: Average Percent Correct on Regular versus PSI Items by Country for the eTIMSS Mathematics Items Selected for eTIMSS 2019 Data Collection



At both grades, the points for the country averages fit closely to the linear regression line, indicating that countries' average percent correct on the items within the PSIs was highly consistent with performance on the regular mathematics items. The correlation between these two percent correct scores was also strong, positive, and statistically significant at both grades (Grade 4: $r(31) = 0.97, p < 0.001$; Grade 8: $r(22) =$

0.96, $p < 0.001$). The high degree of consistency across countries' performance on the two item types provides evidence that even though the PSI items in the field test were more difficult, the regular and PSI mathematics items measured the same construct.

Underlying Factor Structure

The relative model fit of the series of CFA and IRT models used to investigate the underlying factor structure of the assessment also provided evidence that the regular and PSI items are a unidimensional construct. The results from the two analysis approaches used—aggregating the data to the class level, then including students' responses to science items in the models fit with the student-level data—produced slightly different views of the underlying factor structure, but supported the same conclusion.

Exhibit 4.15 presents the number of parameters, dimensions, deviance, AIC, and BIC for the four models fit at each grade with the aggregated class-level data.

Exhibit 4.15: Number of Parameters, Dimensions, Deviance, AIC, and BIC for Class-Level Analysis Models

Class-level Models	Par	Dim	Deviance	AIC	BIC
Grade 4					
Model A: Unidimensional	309	1	104068	104686	106441
Model B: Two-dimensional, uncorrelated	309	2	103571	104189	105944
Model C: Two-dimensional, correlated	310	2	101911	102531	104291
Model D: Bi-factor	412	3	98885	99709	102049
Grade 8					
Model A: Unidimensional	324	1	22760	23408	25108
Model B: Two-dimensional, uncorrelated	324	2	22695	23343	25043
Model C: Two-dimensional, correlated	325	2	21099	21749	23454
Model D: Bi-factor	432	3	17851	18715	20981

Par = number of parameters; Dim = number of dimensions; Deviance = $-2 \times \log\text{-likelihood}$;

AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.

Diagrams of all models are provided in Exhibit 3.11.

Based on the AIC and BIC values, the bi-factor models (Model D) provided the best fit at both grades. The bi-factor models had substantially lower AIC and BIC than

the second best fitting model, the two-dimensional model with correlated factors (Model C—Grade 4: Δ AIC = 2822, Δ BIC = 2242; Grade 8: Δ AIC = 3034, Δ BIC = 2473) as well as the unidimensional model, which was considered the baseline model in this analysis (Model A—Grade 4: Δ AIC = 4977, Δ BIC = 4392; Grade 8: Δ AIC = 4693; Δ BIC = 4127). At both grades, these results suggest that there is some unique common variance specific to the regular and PSI items beyond the general factor, but that the two groups of items are more similar than different. Modeling the two groups of items as specific factors under a general dimension (Model D) rather than completely independent factors (Model C) provided better fit to the data. Further, the correlations between the independent factors in Model C were high (Model C—Grade 4: $r = 0.80$; Grade 8: $r = 0.87$), indicating that treating the groups of items as independent factors was unnecessary.

Although the bi-factor models (Model D) had better fit than the unidimensional models (Model A), it is only appropriate to use this more complex model when the items' have meaningful relationships with the specific factors as well. The median standardized factor loadings for each model were compared to further investigate the differences between the regular and PSI items' relationships with mathematics ability and the importance of the specific factors. Exhibit 4.16 shows the median standardized factor loadings by item type on the general and specific factors for each fourth and eighth grade model fit with the class-level data.

At both grades, the median standardized factor loadings for the regular items on the general factors were moderate and remained highly consistent across the four models fit, regardless of whether this factor was TIMSS mathematics (Models A and D) or

regular mathematics (Models B and C). The median loadings for the regular fourth grade items were approximately 0.46 and the median loadings for the regular eighth grade items were approximately 0.59, indicating that the regular items at both grades had a meaningful relationship with mathematics ability.

Exhibit 4.16: Median Standardized Factor Loadings for Class-Level Analysis Models

Class-level Models	Median Standardized Factor Loadings			
	General Factors		Specific Factors	
	Regular Items	PSI Items	Regular Items	PSI Items
Grade 4				
Model A: Unidimensional	0.45	0.45	—	—
Model B: Two-dimensional, uncorrelated	0.46	0.51	—	—
Model C: Two-dimensional, correlated	0.46	0.51	—	—
Model D: Bi-factor	0.46	0.41	−0.04	0.29
Grade 8				
Model A: Unidimensional	0.59	0.56	—	—
Model B: Two-dimensional, uncorrelated	0.60	0.62	—	—
Model C: Two-dimensional, correlated	0.60	0.62	—	—
Model D: Bi-factor	0.58	0.54	0.14	0.12

In Models A and D all items load on the same general factor—TIMSS mathematics. In Models B and C, there are two general factors—Regular mathematics and PSI mathematics. Diagrams of all models are provided in Exhibit 3.11.

The magnitude of the median standardized factor loadings for the PSI items on the general factors varied slightly across the models fit, but were very consistent with the regular items. In the unidimensional models (Model A), the PSI items had approximately the same median loading on the TIMSS mathematics factor as the regular items at both grades (Model A—Grade 4: $\lambda_{median} = 0.45$; Grade 8: $\lambda_{median} = 0.56$), providing strong evidence that the PSI items are measuring the same mathematics ability as the regular items. The median standardized factor loadings for the PSI items were slightly higher in the two-dimensional models in which the PSI items comprised their own factor (Models B and C—Grade 4: $\lambda_{Gmedian} = 0.51$; Grade 8: $\lambda_{Gmedian} = 0.62$). However, because these

median factor loadings were not substantially different than those seen under the unidimensional model, it was concluded that a model with a single factor for mathematics ability (either Model A or Model D) was a better representation of the relationships among the items.

In the fourth grade bi-factor model, the median factor loading of the PSI items on the specific factor was 0.29, which suggests that the specific PSI factor in this model is helping to explain a small amount of variance among the PSI items that is not accounted for by the general factor. However, the median factor loading for the regular fourth grade items on the specific factor was slightly negative ($\lambda_{REGmedian} = -0.04$), indicating that the specific factor is completely unwarranted for the regular fourth grade items. In the eighth grade bi-factor model, the median loadings on the specific factors for both groups of items were small and could also be considered inconsequential (Model D—Grade 8: $\lambda_{REGmedian} = 0.14$; $\lambda_{PSImedian} = 0.12$).

Although the model fit indices had indicated that the bi-factor models provide better fit than the unidimensional models, the magnitude of the median factor loadings on the specific factors at both grades suggests that these factors are not meaningful beyond the general factor. Therefore, the results obtained with the class-level data at both grades provide evidence that the regular and PSI items can be considered a unidimensional construct.

The unidimensional models and bi-factor models were fit for a second time using the student-level data including the science items to corroborate the results obtained with the class-level data. The student-level results are assumed to provide a more accurate

view of the underlying factor structure because these models were fit to the data in its original format and there is more variability in the item responses at the student-level.

Exhibits 4.17 and 4.18, respectively, present the model fit indices and median

standardized factor loadings for Model A2 and Model D2 fit with the student-level data.

Exhibit 4.17: Number of Parameters, Dimensions, Deviance, AIC, and BIC for Student-Level Analysis Models

Student-level Models	Par	Dim	Deviance	AIC	BIC
Grade 4					
Model A2: Unidimensional	407	2	1495070	1495884	1499415
Model D2: Bi-factor	510	4	1491445	1492465	1496890
Grade 8					
Model A2: Unidimensional	473	2	1174081	1175027	1178975
Model D2: Bi-factor	432	3	1167511	1168671	1173512

Par = number of parameters; Dim = number of dimensions; Deviance = $-2 \times \log\text{-likelihood}$;

AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.

Diagrams of all models are provided in Exhibit 3.12.

Exhibit 4.18: Median Standardized Factor Loadings for Student-Level Analysis Models

Student-level Models	Median Standardized Factor Loadings			
	General Factors		Specific Factors	
	Regular Items	PSI Items	Regular Items	PSI Items
Grade 4				
Model A2: Unidimensional	0.56	0.61	—	—
Model D2: Bi-factor	0.55	0.59	0.08	0.13
Grade 8				
Model A2: Unidimensional	0.66	0.75	—	—
Model D2: Bi-factor	0.67	0.70	0.01	0.08

In Models A and D all items load on the same general factor—TIMSS mathematics.

Diagrams of all models are provided in Exhibit 3.12.

At both grades, the results obtained with the student-level data were highly consistent with those obtained with the class-level data. Based on the AIC and BIC values, the student-level bi-factor models also provided superior fit compared to the unidimensional models (Model A2—Grade 4: $\Delta\text{AIC} = 3419$, $\Delta\text{BIC} = 2525$; Grade 8: $\Delta\text{AIC} = 6356$, $\Delta\text{BIC} = 5463$). The median standardized factor loadings on the general

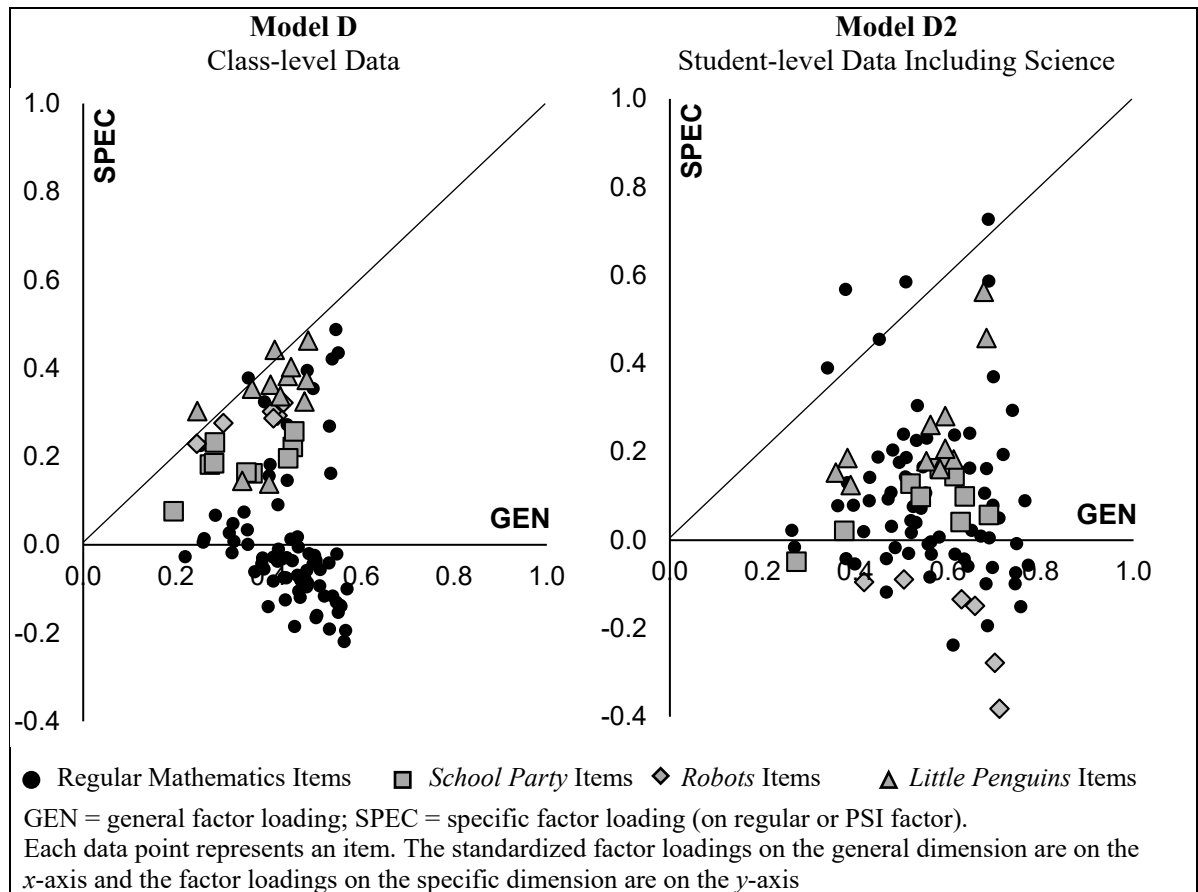
factors for both the regular and PSI items at both grades were higher than those seen in the models fit with the class-level data, providing further evidence of the meaningful relationship between both groups of items and mathematics ability. Further, compared to the class-level bi-factor models, the median loadings on the specific factors were even smaller, providing stronger evidence in favor of the unidimensional models.

Plots of the items' factor loadings in the bi-factor models were used to confirm that the specific factors in these models were not warranted. Exhibit 4.19 shows plots of the fourth grade items' factor loadings for the bi-factors models fit using both the class-level data and student-level data including science. Each round black marker represents a regular item and each grey marker represents a PSI item. The shapes of the markers for the PSI items denote the tasks to which the items belong. The factor loadings on the general factor are on the x -axis and the factor loadings on the specific factor to which each item is assigned are on the y -axis. The identity line ($y = x$) shows the line below which all the items will appear if their loading on the general dimension was greater than their loading on the specific dimension. Any items with higher loadings on the specific factor (i.e., appearing above the identity line) may not be contributing to the general construct.

On the plot produced with class-level data (left), the regular items have slightly higher loadings on the general factor than the PSI items. The PSI items all have positive loadings on their specific factor and are clustered higher up on the y -axis close to the identity line, showing that the specific PSI factor is accounting for some unique variance beyond the general factor. However, the majority of the regular items are clustered

around and below the x -axis, showing that the regular items are unrelated to or have a negative relationship with their specific factor. This means that the regular items did not have any unique shared variance to explain beyond the general factor. There is also some clustering among the PSI items by task, which may be related to the shared context or because the items in most tasks address the same content domain.

Exhibit 4.19: Plots of Standardized Factor Loadings on General versus Specific Factors from Bi-factor Models – Grade 4



In the fourth grade plot produced with the student-level data (right), both item types are clustered around the x -axis and there are some regular and PSI items with negative loadings on their specific factors. This suggests that the specific factors are not helping to explain a substantial amount of unique variance among either group of items

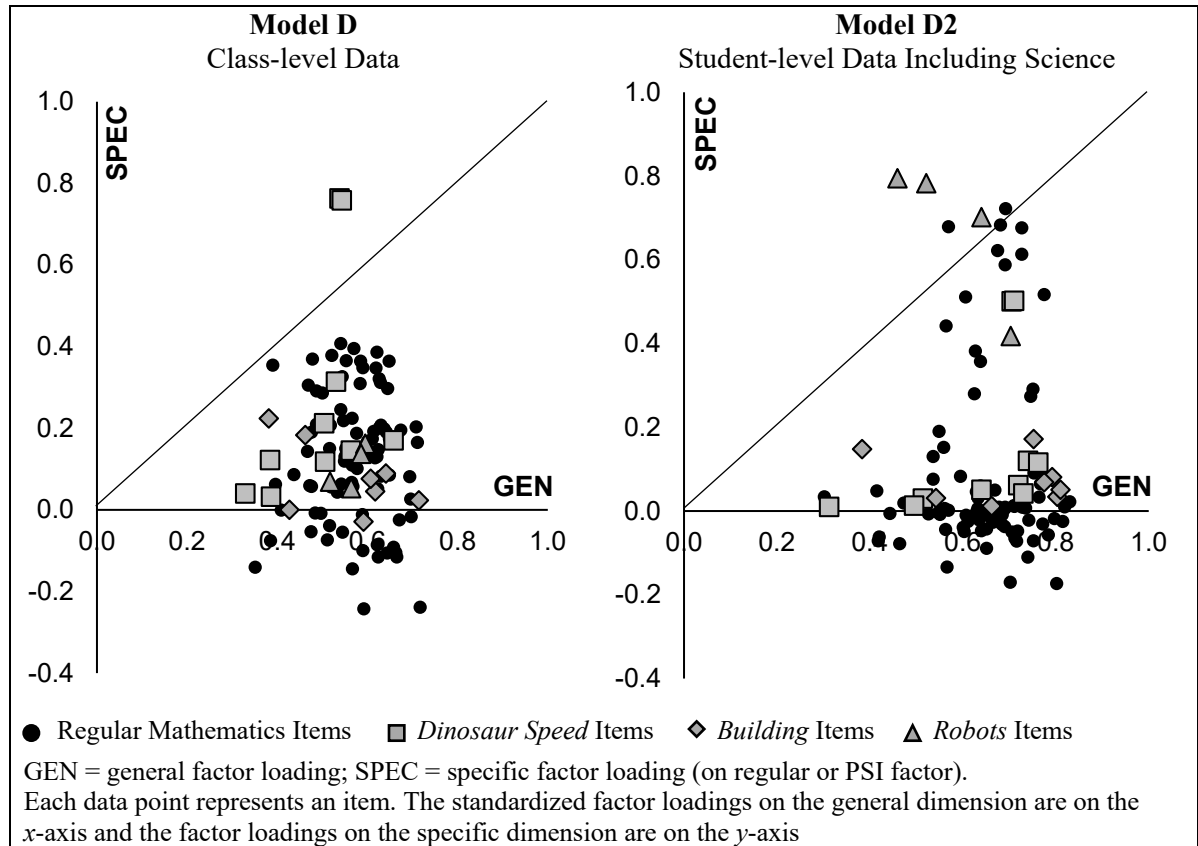
beyond the general factor. Several outlying items have moderate to high loadings on their specific factors, including five regular items that fall above the identity line, but these items appear to be anomalies in an otherwise clear cluster of items around the x -axis.

Both of the fourth grade plots support the conclusion that the fourth grade mathematics items can be considered a unidimensional construct. Although the PSI items had slightly higher loadings on their specific factor in the class-level plot, it is suspected that these higher loadings were only seen with the class-level data because contextual variables such as teachers' use of technology in the classroom and emphasis on problem solving in mathematics may have a greater impact on classes' scores across the two item types than on individual students' scores.

Exhibit 4.20 shows the parallel plots for the eighth grade bi-factor models. At the eighth grade there are some differences in the clustering of the items across the two plots, but they both present the same general view of the underlying factor structure of the eighth grade mathematics assessment. In the class-level plot (left), the regular and PSI items are all clustered together with similar loadings on both the general factor and their respective specific factors. This shows that the specific factors are accounting for about the same small amount of unique shared variance among both groups of items. With the exception of two outlying PSI items, all of the eighth grade mathematics items fall below the identity line, showing that both item types are more representative of the general mathematics factor than their respective specific factors. The outlying items in the class-level plot are two parts of a single item that address the same framework topic, which may explain why these items have very similar loadings on both factors. However, there

is no clear explanation for why these items depended so highly on the specific factor in comparison to the other items.

Exhibit 4.20: Plots of Standardized Factor Loadings on General versus Specific Factors from Bi-factor Models – Grade 8



The eighth grade plot produced with the student-level data including the science items (right) shows a similar pattern, but there are some notable differences. The majority of the items are in an even tighter cluster around the x -axis than was seen with the class-level data, but there are a smaller number of more dispersed items with stronger loadings on their respective specific factors, including several items that fall above the identity line. None of the eighth grade PSI items have negative loadings on the specific PSI factor, although most are close to zero. The PSI items that fall above the identity line are

all part of the same task, *Robots*, which includes several interactive items designed to assess students' reasoning skills in algebraic relationships and functions.

Taken all together, the results show that the unidimensional model provides a good representation of the combined regular and PSI mathematics items at both grades, and that these items can be validly reported as a single mathematics construct. Although there is some evidence that the bi-factor model provides slightly better fit to the data, it is insufficient to compensate for the added complexity in analysis and reporting.

Summary of Results

Taken together, the validity evidence based on test content, response process, and internal structure presented in this dissertation provides evidence that both the fourth and eighth grade mathematics PSIs deliver valid measurement of students' mathematics ability as defined by TIMSS. At both grades, the investigation into the structural relationship between the PSIs and regular mathematics items suggests that the PSIs can be reported together with the regular eTIMSS mathematics items.

The rigorous methods used in developing the mathematics PSIs ensured that the tasks were aligned with the mathematics content and skills outlined in the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017) and well-suited for TIMSS' diverse target population. These methods also provided evidence that the PSIs met the first goal for the tasks—assess mathematics problem solving. Adding the mathematics PSIs to the pool of regular items in the eTIMSS 2019 mathematics assessments resulted in a small increase in coverage of the applying and reasoning cognitive domains at both the fourth and eighth grade as intended, but did not lead to substantial deviations from the target

percentages of testing time allocated to each domain in the framework. Therefore, the PSI items can be included in the eTIMSS 2019 achievement scales without skewing the assessments' coverage of the framework.

A thorough investigation of the available sources of qualitative and quantitative data collected for the eTIMSS 2019 Field Test provided evidence in support of the response process validity of the eTIMSS assessments and specifically the mathematics PSIs. Being the first large-scale administration of TIMSS on computers and tablets, the eTIMSS 2019 Field Test also was a critical “dress rehearsal.” It prompted a number of improvements to the eTIMSS assessment systems, directions, and PSIs for main data collection that are expected to further enhance the response process validity of the assessments.

Overall, the eTIMSS Player was reliable in delivering the eTIMSS 2019 Field Test to students and enabled students to navigate through the achievement items with ease. The eTIMSS field test instruments, and particularly the PSIs, included a wide variety of item types that were generally well-received by students. A comparison of the field test item statistics across the digital and paper forms of the field test instruments as well as observations of students' interaction with the enhanced item types provided evidence that the eTIMSS item types largely elicited the intended interactions from students.

Students also found eTIMSS to be engaging. At both grades, the majority of students reported that they enjoyed taking eTIMSS on a digital device. Analyses of the timing data collected during the field test indicated that students' time on task increased

with the cognitive demand of the items, providing support for the validity of the cognitive domain classifications assigned to the items as well as further evidence of students' effort and motivation in taking the test.

TIMSS efforts to ensure valid and reliable scoring of both human- and machine-scored constructed response items were successful. Both the regular and PSI items that required human scoring via the IEA's Online Scoring System were scored with a high degree of reliability. Also, the scores assigned to all machine-scored items were verified.

Analyses of the psychometric properties and underlying structure of the regular and PSI items provided evidence that the PSI items are measuring the same mathematics ability as the regular eTIMSS items. At both grades, the PSI items in the field test were more difficult than the regular mathematics items, but countries' average percent correct scores across the two groups of items were highly correlated. The PSI items were equally as highly discriminating as the regular items, providing evidence that the PSIs were successful in differentiating between high and low achieving students in traditionally challenging to measure areas of the framework. Finally, fitting a series of confirmatory factor analysis models to the data provided evidence that the regular and PSI items at both the fourth and eighth grades can be considered a unidimensional construct and scaled together.

Chapter 5: Discussion

TIMSS expended significant effort and resources developing mathematics PSIs at the fourth and eighth grades for eTIMSS 2019 with the goal of improving measurement of students' mathematics problem solving skills and enhancing the validity of the TIMSS 2019 achievement scales. The mathematics PSIs were developed to measure the same *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017) as the regular eTIMSS items, but were a unique and somewhat experimental addition to the eTIMSS 2019 assessments. The items within the mathematics PSIs were situated within extended problem contexts, placed more emphasis on higher-order thinking skills, and included more interactive features and enhanced item types than the regular eTIMSS mathematics items. Therefore, although the PSI items were designed to measure the same underlying construct as the regular eTIMSS items, there was a question about whether these novel tasks extended the fourth and eighth grade eTIMSS 2019 mathematics assessments as intended, or measured a different construct. It was important to address this question before TIMSS reported the results of the eTIMSS 2019 assessments.

The primary purpose of this dissertation was to conduct an in-depth investigation of the validity of the eTIMSS 2019 mathematics PSIs to inform decisions about analyzing and reporting the results for the PSIs in 2019 as well as the future of the tasks in subsequent TIMSS assessment cycles. TIMSS needed to decide if the items within the PSIs were similar enough to the regular fourth and eighth grade mathematics items to be reported on the TIMSS 2019 achievement scales or were a different construct, requiring separate analysis and reporting. Also, because developing PSIs is a highly resource

intensive process, TIMSS needed to determine if the PSIs were worth the effort and should be integrated into future assessment cycles, beginning with the next assessment in 2023.

The second purpose of this dissertation was to contribute to the growing body of literature surrounding technology-enhanced assessment. TIMSS learned a number of valuable lessons in developing the PSIs which are important to document and share to support continued progress toward realizing the full potential of digital assessments in general and mathematics assessment in particular.

Three key tasks were completed to meet these dissertation goals:

- 1) Examining and documenting the methods and procedures used to develop the PSIs;
- 2) Investigating the characteristics of the PSIs in terms of content coverage and fidelity of student responses; and
- 3) Using the eTIMSS field test data to evaluate the internal structure of the PSIs.

Chapter 3 documented the substantial undertaking of developing the mathematics PSIs for eTIMSS 2019. The nearly four-year development process required close collaboration among staff at the TIMSS & PIRLS International Study Center, a staunch group of expert mathematics consultants, and software developers and programmers at IEA Hamburg to ensure that the PSIs came to fruition and were presented to students as intended. In addition to countless rounds of iterative review involving mathematics experts and country representatives, a series of cognitive laboratories and pilot tests were conducted to guide decisions about the mathematics content and eTIMSS user interface.

Chapter 4 further addressed the test content validity of the mathematics PSIs by showing how perfectly the eTIMSS 2019 mathematics PSIs aligned with the framework as well as improved the framework coverage provided by the eTIMSS 2019 mathematics assessments.

Chapter 4 also used the data collected in the eTIMSS 2019 Field Test in over 30 countries to investigate whether students interacted with the mathematics PSIs as intended and how the PSI items fit in with the regular eTIMSS mathematics items. Response process validity was examined using data that showed students had little difficulty interacting with the PSIs as well as how much the students liked and engaged with the PSIs. Also, the scoring reliability of the constructed response items within the PSIs was examined to ensure that students' responses to the items within the tasks were scored with a high degree of reliability (97%). The structural relationship between the PSI items and the regular eTIMSS items was evaluated by comparing the psychometric properties and countries' performance across the two groups of items and fitting a series of confirmatory factor analysis models to the data to investigate the underlying relationships among the regular and PSI items. These results indicated that the PSI items measured the same mathematics ability as the regular mathematics items.

Summary of Key Findings

The evidence presented in this dissertation indicates that both the fourth and eighth grade mathematics PSIs deliver valid measurement of the same mathematics ability as the regular eTIMSS mathematics items. The PSIs are aligned with the mathematics framework and enhanced coverage of the mathematics applying and

reasoning cognitive domains without skewing the amount of testing time allocated to each content and cognitive domain. Therefore, from a content perspective, it is appropriate to include the PSIs in the eTIMSS 2019 achievement scales.

The evidence of response process validity gathered in the eTIMSS 2019 Field Test indicates that students generally interacted with the mathematics PSIs as intended and found eTIMSS to be engaging and motivating, suggesting that the PSIs and eTIMSS assessments elicited the intended responses from students. Also, it was confirmed that both the machine- and human-scored PSI items in the eTIMSS 2019 Field Test were scored reliably, supplying further evidence that scores on the PSI items can be considered valid.

The analyses of the psychometric properties and underlying structure of the regular and PSI items conducted with the eTIMSS 2019 Field Test data provided robust evidence that the PSI items are measuring the same mathematics ability as the regular eTIMSS items. At both grades, countries' average percent correct scores across the two groups of items were highly correlated. The PSI items were more challenging than the regular items, but they were equally as highly discriminating, demonstrating that the PSIs were successful in enhancing measurement of traditionally challenging to measure areas of the framework. The unidimensional model provided a good representation of the combined regular and PSI mathematics items at both grades, indicating that the two item types can be validly reported on the same scale.

Lessons Learned

In the early stages of the PSI development process, the Executive Directors of the TIMSS & PIRLS International Study Center established four criteria for a successful mathematics PSI—1) assess mathematics problem solving; 2) take advantage of the “e” environment; 3) be engaging and motivating for students; and 4) be administered and scored via the TIMSS eAssessment systems. It was expected that meeting these criteria would be a challenge for TIMSS, especially because 2019 was TIMSS’ inaugural cycle as a digital assessment, but the actual scope of this undertaking exceeded the initial expectations.

Everyone involved in the PSI development process, including the author, learned an immense amount about digital assessment of mathematics problem solving, writing coherent item sets, leveraging technology to support valid measurement, and working with developing software throughout the development process. To provide a comprehensive summary of the most important lessons learned, the author asked the group of individuals most involved in the eTIMSS 2019 mathematics PSI development work to reflect on what they learned and could now recommend for future assessments. This group included the Executive Directors of the TIMSS & PIRLS International Study Center, the Director of User Interface and Software Development at the TIMSS & PIRLS International Study, and the two mathematics education and measurement experts that provided the ideas for the PSIs and participated in the series of meetings with TIMSS staff and other experts to develop and refine the tasks. Considering these multiple perspectives, two overarching lessons were articulated.

Characteristics of Successful Extended Mathematics Problem Contexts

Extended assessment tasks are becoming more and more common with the rise of electronic educational assessment. However, the majority of the progress made in this direction thus far has been in subject areas with well-established histories of practical investigation (e.g., the sciences) or for 21st century skills that arose in direct response to the current computer age. In mathematics, digital technology is increasingly being used in educational games or as a mechanism to provide immediate feedback to students on problem sets, but there are still few examples of extended, real world, assessment tasks in mathematics. Therefore, without a strong tradition in this method of assessment in mathematics, developing the mathematics PSIs required innovating in largely uncharted territory.

The development goals for the mathematics PSIs included many competing demands that proved to be extremely difficult to satisfy at the same time. It was a major challenge to devise problem contexts that were suitable for addressing the *TIMSS 2019 Mathematics Framework* (Lindquist et al., 2017), could be investigated through a series of independent items, were interesting and engaging for students, and supported valid uses of technology. Throughout the development process, TIMSS learned more about the characteristics of a successful mathematics problem scenario for a large-scale international assessment.

A successful problem context needs to be complex enough to warrant extended mathematical investigation, but not so complex that it requires lengthy explanation or the introduction of technical jargon. Ideally, a problem context is authentic, but not to the

point that the authenticity detracts from the mathematics at hand. From the beginning, it is essential to consider the difficulty level of the items the context may support, as many interesting real-world problems cannot be simplified to fourth or eighth grade level mathematics. Particularly at the fourth grade, more structured scenarios were found to be better suited for assessing the mathematics framework and were therefore selected over more realistic tasks. Contexts that can be used to address a range of content domain topics and cognitive processes also are preferable to those that focus on a single area to ensure even coverage of the framework. However, creating problem contexts that span multiple content domains proved to be more challenging than contexts that focused on one or two domains.

A problem context must be interesting and grade-appropriate so that it is engaging and motivating for students. At the same time, it cannot be too new to any students because this could introduce cultural bias that can compromise the content validity of the assessment. Particularly in an international context, this requirement is very challenging to meet because students around the world come from many different cultures and experiences. Although this criteria is important for all achievement items, it becomes even more vital when the context will span a series of items.

It is essential that a problem context can be studied through a series of independent items. However, most real-world applications of mathematics involve a progression of interdependent steps, which are challenging to adapt for use in an assessment situation when a single misstep can limit students' opportunities to demonstrate their ability on the rest of the items. The most successful mathematics

contexts were found to be overarching problems with a series of sub-tasks to be completed. For example, if the overarching task is to plan a school event, independent questions can be asked about the food, drinks, and decorations to buy given various circumstances. However, introducing conditions that span the entire task (e.g., a total budget for the event) typically leads to complicated and unnecessary dependencies among the items that should be avoided. Also, none of the answers to questions in the task should be given away by information on another screen. Although it is possible to track if students changed their answers based on a clue on another screen, it becomes controversial to determine which answer should be considered in scoring.

Finally, it is important to consider whether a problem context lends itself to valid uses of technology. The PSIs were conceived of as a way to further capitalize on the benefits of eAssessment, so contexts that can benefit from interactive stimuli, animations, or more complex response spaces are best suited for PSIs. When a context does not require such features, the authors must be careful to avoid superfluous uses of technology that will only distract from the mathematics.

Challenges of Developing Software and Content in Tandem

Because eTIMSS 2019 was TIMSS' first cycle as a digital assessment, the development of the eAssessment systems and features coincided with the development of the PSIs. Setting out to transition from creating paper-based test booklets to programming complex digital tasks that imitate the real world and assess skills in more sophisticated ways (i.e., early third generation assessment; Bennett, 2015) was a huge advance from a technical perspective as well. Test content development and software development are

both highly iterative and resource intensive processes that become more difficult when undertaken in tandem, presenting even greater challenges in developing the PSIs.

Developing valid achievement items requires many rounds of review and revision, particularly when aiming to create tasks as complex and unique as the PSIs. Initially, the content of digital items can be drafted and reviewed on paper, but eventually it is necessary to program the tasks to accurately appraise the test content and try out the tasks with students to investigate their interactions with functional versions of the instruments. However, individually programming PSIs within a concurrently evolving eAssessment system was an arduous process in and of itself that also required substantial time and resources. Ultimately, it was not possible or efficient to constantly re-program complex technology-enhanced items. Navigating this tension between the demands of content development and software development is difficult, and attempting both at the same time should be avoided.

Suggestions for the Future

Several suggestions for future TIMSS assessments and other assessment programs seeking to capitalize on technology can be made based on experiences in developing the eTIMSS 2019 mathematics PSIs.

Always consider the target construct first. First and foremost, it is essential to ensure that the target construct that the PSI is designed to measure is the driving force behind all decisions made throughout its development process. When designing complex assessment tasks with many competing demands it becomes exponentially more challenging to maintain this focus. Although including interactive features that capitalize

on the digital environment and promote student engagement and motivation are important, it is critical to ensure that these goals do not detract from the validity of the PSI. Throughout item development it is vital to carefully appraise the purpose and functionality of each use of technology to ensure that it supports measurement of the target construct and is intuitive to use, so that it does not become a distraction.

Typically, students do not need long explanations about the digital features.

Incorporating new technology into an assessment naturally warrants the addition of more directions for students to explain how to interact with the test content. Although it is essential that all students are well prepared to respond to the items, it also is very important to avoid including lengthy explanations of features that students will naturally intuit how to use. Overly detailed directions increase the total time needed to administer the assessment as well as the reading load, which can cause students to lose focus or become fatigued more quickly. All directions should be as short and to the point as possible and written with the understanding that many students today frequently use computers both in and out of the classroom. If the enhanced features used in the assessment are relevant and well-designed, long explanations of how to use them should not be needed.

Keep the reading load to a minimum. Items situated in a cohesive problem solving context may require more reading than standalone questions, but it is essential that the reading load is kept to a minimum. Early drafts of the mathematics PSIs included considerably more text than the regular TIMSS items and after each pilot the text was shortened until the reading load was eventually brought into line with the rest of the

assessment. Though the context is important, only details relevant to the mathematics questions of interest should be included.

Consider all features of the task when determining the testing time. Beyond the number of items in the task, it is important to consider the cognitive demand of the items, reading load, number of interactive features, and the novelty of these features in determining the time needed to complete the task. Despite having a similar number of items as the regular eTIMSS item blocks, the mathematics PSIs took students more time to complete, which in retrospect is not surprising given their complexity. Not providing students with enough time to comfortably complete the tasks will result in unintended speededness and position effects that can impact the measurement properties of the items.

Conduct cognitive laboratories and pilot tests. Pilot testing is always important in developing valid achievement instruments, but it is even more critical to the development process when transitioning to digital assessment and pioneering new item formats. Both small- and large-scale outings are useful and should be conducted as often as possible, particularly in the early stages of development. The series of cognitive laboratories and pilot tests conducted throughout the eTIMSS development process played a vital role in guiding decisions made about the problem contexts, items, uses of technology, time to complete the tasks, user interface, and directions. Each outing brought about substantial improvements that would not have been possible without trying out the tasks in the field.

Do not begin programming too soon. To the extent possible, the mathematics content and uses of technology should be laid out on paper and critically reviewed by

software developers early in the development process. Programming work is costly and time consuming so it is advantageous to invest more time in deciding the specifications for the tasks before investing substantial resources in actualizing programming. It is helpful to involve programmers in the early stages of development to determine what is technically possible and establish a shared understanding of how the drafted content will appear on screen. Programming the tasks will unavoidably result in more questions about how features should function and further external reviews will result in more revisions that are difficult to predict beforehand. Still, having a solid foundational agreement between the content specialists, measurement experts, and software developers about the entire process is key.

Implications for the Future of the PSIs

Developing the eTIMSS 2019 mathematics PSIs was an extremely challenging and resource intensive process, particularly because it coincided with TIMSS' initial transition to digital assessment. However, the resulting PSIs were an important addition to the eTIMSS 2019 assessments and the field of technology-enhanced mathematics assessment. The PSIs addressed applied mathematics problem solving and reasoning skills that are highly valued by the international mathematics education community, but rarely assessed in large-scale studies. By leveraging the digital mode of administration, TIMSS was successful in creating tasks that motivate and guide students through challenging series of mathematics problems that could not have been assessed on paper. Further, the results of this dissertation, indicate that the PSI items can be scaled and reported together with the regular eTIMSS items as intended.

Now that TIMSS has built the foundation of the eTIMSS infrastructure and learned an immense amount about developing successful problem contexts and leveraging technology to support valid measurement, TIMSS is well positioned to continue advancing measurement in this progressive direction. Given the many positive benefits of the PSIs and TIMSS' initial success in this initiative, the PSIs should continue to be a part of the eTIMSS 2023 mathematics assessments and beyond.

References

- Ackerman, T.A., Gierl, M.J. & Walker, C.M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22, 37–51.
- Agencia de Calidad de la Educación (2016). Chile. In Mullis, I.V.S., Martin, M.O., Goh, S. & Cotter, K.E. (Eds.), *TIMSS 2015 Encyclopedia: Education Policy and Curriculum in Mathematics and Science*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website:
<http://timssandpirls.bc.edu/timss2015/encyclopedia/>.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & F. Csáki (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akadémiai Kaidó.
- American Psychological Association, American Educational Research Association & National Council on Measurement in Education [APA, AERA & NCME] (2014). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.
- American Institutes for Research [AIR]. (2015, September). *eTIMSS cognitive interview report*. Prepared for the TIMSS & PIRLS International Study Center at Boston College. Washington, DC: Author.
- Archbald, D. & Newmann, F.M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school*. Reston, VA: National Association of Secondary School Principals.
- Auewarakul, C., Downing, S.M., Jaturatamrong, U. & Praditsuwan, R. (2005). Sources of validity evidence for an internal medicine student evaluation system: An evaluative study of assessment methods. *Medical Education*, 39(3), 276–283.
- Barnes, M., Clarke, D. & Stephens, M. (2010). Assessment: The engine of systemic curricular reform?, *Journal of Curriculum Studies*, 32(5), 623–650.
- Bennett, R.E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing*. Princeton, NJ: Policy Information Center, Educational Testing Service. Retrieved from:
https://www.ets.org/research/policy_research_reports/pic-reinvent.

- Bennett, R.E. (2014). Preparing for the future: What educational assessment must do. *Teachers College Record*, 116(11).
- Bennett, R.E. (2015). The changing nature of educational assessment. *Review of Research in Education*, 39, 370–407. Princeton, NJ: Educational Testing Services.
- Bennett, R.E., Persky, H., Weiss, A. & Jenkins, F. (2003). Assessing complex problem solving performances. *Assessment in Education: Principles, Policy & Practice*, 10(3), 347–359.
- Bennett, R.E., Persky, H., Weiss, A. & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP technology-based assessment project, research and development series*. U.S. Department of Education Institute of Education Sciences.
- Bennett, R.E., Persky, H., Weiss, A. & Jenkins, F. (2010). Measuring problem solving with technology: A demonstration study for NAEP. *The Journal of Technology, Learning and Assessment*, 8(8).
- Birenbaum, M. & Tatsuoaka, K.K. (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. *Applied Psychological Measurement*, 11(4), 385–395.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical Theories of Mental Test Scores* (pp. 397–472). Reading, MA: Addison-Wesley Publishing.
- Boaler, J. (1993). The role of contexts in the mathematics classroom: do they make mathematics more 'Real'?. *For the Learning of Mathematics*, 13(2), 12–17.
- Boaler, J. & Staples, M. (2008). Creating mathematical futures through an equitable teaching approach: The case of railside school. *Teachers College Record*, 110(3), 608–645.
- Bradow, E.T., Wainer, H. & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64, 153–168.
- Bradow, E.T., Wainer, H. & Wang, X. (Eds.) (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.
- Braun, H. (2013). Prospects for the future: a framework and discussion of directions for the next generation of international large-scale assessments. In M. Von Davier, E. Gonzalez & I. Kirsch (Eds.), *The Role of International Large-Scale Assessments*:

- Perspectives from Technology, Economy, and Educational Research* (pp. 149–160).
- Brown, T.A. (2014). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Publications. Retrieved from ProQuest Ebook Central.
- Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research and Evaluation*, 22(1).
- Burnham, K.P. & Anderson, D.R. (2002). Model selection and multimodel inference: A practical information-theoretic approach (2nd ed.). New York, NY: Springer.
- Cepeda, N.J., Blackwell, K.A. & Munakata, Y. (2013). Speed isn't everything: Complex processing speed measures mask individual differences and developmental changes in executive control. *Developmental Science*, 16, 269–286.
- Chen, F.F., West, S.G., & Sousa, K.H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, 41, 189–225.
- Cizek, G.J., Rosenberg, S.L. & Koons, H.H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, 68, 397–412.
- Cook, D.A., Zendejas, B., Hamstra, S.J., Hatala, R. & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in Health Sciences Education*, 19(2), 233–250.
- Cormier, D.C., Yeo, S., Christ, T.J., Offrey, L.D. & Pratt, K. (2016). An examination of the relationship between computation, problem solving, and reading. *Exceptionality*, 24(4), 225–240.
- Crespo, S. & Sinclair, N. (2008). What makes a problem mathematically interesting? Inviting prospective teachers to pose better problems. *Journal of Mathematics Teacher Education*, 11, 395–415.
- Cucina, J. & Byle, K. (2017). The bifactor model fits better than the higher-order model in more than 90% of comparisons for mental abilities test batteries. *Journal of Intelligence*, 5(3).
- Darling-Hammond, L., Herman, J., Pellegrino, J. Abedi, J., Aber, J.L., Baker, E. ...Steele, C.M. (2013). Criteria for high-quality assessment. SCOPE, CRESST, LSRI Policy Brief. Retrieved from: <http://edpolicy.stanford.edu>.

- Darling-Hammond, L. & Lieberman, A. (1992). *The shortcomings of standardized tests*. The Chronicle of Higher Education, B1–B2.
- Davis, R.B. (1992). Reflections on where mathematics education now stands and on where it may be going. In D.A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 724–734). New York, NY: Macmillan.
- de Ayala, R. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.
- de la Torre, J. & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, 33(8), 620–639.
- DeMars, C.E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, 43, 145–168.
- DeMars, C.E. (2013). A tutorial on interpreting bi-factor model scores. *International Journal of Testing*, 13, 354–378.
- Desimone, L.M. & Carlson, K.L.F. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis*, 26(1), 1–22.
- Dolan, R.P., Goodman, J., Strain-Seymour, E., Adams, J. & Sethuarman, S. (2011, March). *Cognitive lab evaluation of innovative items in mathematics and english language arts assessment of elementary, middle, and high school students*. Pearson Assessment Research Report.
- Dolan, R.P., Strain-Seymour, E., Way, W.D. & Rose, D.H. (2013, April). *A universal design for learning-based framework for designing accessible technology-enhanced assessments*. Pearson Assessment Research Report.
- Dossey, J.A. (2003). Large-scale assessments: National and international. In G. Stanic & J. Kilpatrick (Eds.), *The History of School Mathematics* (pp. 1435–1490). Reston, VA: The National Council of Teachers of Mathematics.
- Drasgow, F., Luecht, R.M. & Bennett, R.E. (2006). Technology and testing. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Westport, CT: American Council on Education and Praeger.
- Evers, A., Sijtsma, K., Lucassen, W. & Meijer, R.R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing*, 10(4), 295–317.

- Fischer, A., Greiff, S., Wüstenberg, S., Fleischer, J., Buchwald, F. & Funke, J. (2015). Assessing analytic and interactive aspects of problem solving competency. *Learning and Individual Differences*, 39, 172–179.
- 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).
- Foy, P. (2017). *TIMSS 2015 user guide for the international database*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <https://timssandpirls.bc.edu/timss2015/international-database/>.
- Foy, P., Martin, M.O., Mullis, I.V.S., Yin, L., Centurino, V.A.S. & Reynolds, K.A. (2016). Reviewing the TIMSS 2015 achievement item statistics. In M.O. Martin, I.V.S. Mullis & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 11.1–11.43). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <http://timss.bc.edu/publications/timss/2015-methods/chapter-11.html>.
- Funke, J. (2009). Complex problem solving: A case for complex cognition? *Cognitive Processing*, 11(2), 133–142.
- Gibbons, D.R. & Hedeker, D. (1992). Full information item bi-factor analysis. *Psychometrika*, 57, 423–436.
- Gorin, J.S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, 25, 21–35.
- Greatorex, J. (2013). *Context in mathematics examination questions*. Cambridge, MA: Cambridge Assessment, Assessment Research and Development.
- Greiff, S., Holt, D.V. & Funke, J. (2013). Perspectives on problem solving in educational assessment: Analytical, interactive, and collaborative problem solving. *Journal of Problem Solving*, 5(2).
- Greiff, S., Niepel, C., Scherer, R. & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, 61, 36–46.
- Greiff, S., Wüstenberg, S. & Funke, J. (2012). Dynamic problem solving: A new measurement perspective. *Applied Psychological Measurement*, 36(3), 189–213.

- Grønmo, L.S., Lindquist, M., Arora, A. & Mullis, I.V.S. (2013). TIMSS 2015 mathematics framework. In I.V.S. Mullis & M.O. Martin (Eds.), *TIMSS 2015 Assessment Frameworks* (pp. 11–27). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <https://timssandpirls.bc.edu/timss2015/frameworks.html>.
- Gustafsson, J.E., Nilsen, T. & Hansen, K.Y. (2018). School characteristics moderating the relation between student socio-economic status and mathematics achievement in grade 8. Evidence from 50 countries in TIMSS 2011. *Studies in Educational Evaluation*, 57, 16–30.
- Hair, J.F., Black, W.C., Babin, B.J. & Anderson, R.E. (2014). *Multivariate data analysis* (7th ed.). New York, NY: Pearson.
- Herde, C.N., Wüstenberg, S. & Greiff, S. (2016). Assessment of complex problem solving: What we know and what we don't know. *Applied Measurement in Education*, 29(4), 265–277.
- Hiebert, J. & Wearne, D. (1993). Instructional tasks, classroom discourse, and students' learning in second-grade arithmetic. *American Educational Research Journal*, 30(2), 393–425.
- Holzinger, K.J., & Harman, H.H. (1938). Comparison of two factorial analyses. *Psychometrika*, 3, 45–60.
- Holzinger, K.J. & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2, 41–54.
- Hopfenbeck, T.N. & Maul, A. (2011). Examining evidence for the validity of PISA learning strategy scales based on student response processes. *International Journal of Testing*, 11(2), 95–121.
- Huff, K.L. & Sireci, S.G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice*, 20(3), 16–25.
- Huff, K., Steinberg, L. & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education*, 23(4), 310–324.
- Husen, T. (Ed.) (1967). *International study of achievement in mathematics: A comparison of twelve countries*. (Vols. 1 & 2). New York, NY: Wiley.
- International Association for the Evaluation of Educational Achievement [IEA] (2005). *TIMSS special initiative in problem solving and inquiry*. Retrieved from Boston

- College, TIMSS & PIRLS International Study Center website:
<https://timssandpirls.bc.edu/timss2003i/psi.html>.
- International Association for the Evaluation of Education [IEA] (2017). *A brief history of the IEA*. Retrieved from: <http://www.iea.nl/brief-history-iea>.
- Jodoin, M.G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 40(1), 1–15.
- Johansson, S. (2016). International large-scale assessments: What uses, what consequences? *Educational Research*, 58(2), 139–148.
- Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183–202.
- Jöreskog, K.G. (1971a). Simultaneous factor analysis in several populations. *Psychometrika*, 36, 409–426.
- Jöreskog, K.G. (1971b). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36, 109–133.
- Kane, M.T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319–342.
- Kane, M.T. (2006). Validation. In B.L. Robert (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Wesport, CT: Praeger.
- Kaplan, D. (2009). *Advanced quantitative techniques in the social sciences: Structural equation modeling foundations and extensions* (2nd ed., Vols. 1–10). Thousand Oaks, CA: SAGE Publications, Inc.
- Kilpatrick, J. (1992). A history of research in mathematics education. In D.A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 3–38). New York, NY: Macmillian.
- Kind, P.M. (2013). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning. *Journal of Research in Science Teaching*, 50(5), 530–560.
- Klahr, D. & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, 12(1), 1–48.
- Kreiter, C. (2015). When I say ... response process validity. *Medical Education*, 49(3), 247–248.

- Kuo, C.Y., Wu, H.K., Jen, T.H. & Hsu, Y.S. (2015). Development and validation of a multimedia-based assessment of scientific inquiry abilities. *International Journal of Science Education*, 37(14), 2326–2357.
- Kupiainen, S., Vainikainen, M.P., Marjanen, J. & Hautamäki, J. (2014). The role of time on task in computer-based low-stakes assessment of cross-curricular skills, *Journal of Educational Psychology*, 106(3), 627–638.
- LaRoche, S. (2017). *TIMSS 2019 Sampling*. Presentation presented at the 1st TIMSS 2019 NRC Meeting. Hamburg, Germany.
- LaRoche, S., Joncas, M. & Foy, P. (2016). Sample Design in TIMSS 2015. In M.O. Martin, I.V.S. Mullis & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 3.1–3.37). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <http://timss.bc.edu/publications/timss/2015-methods/chapter-3.html>.
- Lee, Y.H. & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, 53, 359–379.
- Liljedahl, P., Santos-Trigo, M., Malaspina, U. & Bruder, R. (2016). Problem solving in mathematics education. In K. Gabriele (Ed.), *ICME-13 Topical Surveys* (pp. 1–38). Springer International Publishing: Springer.
- Lindquist, M., Philpot, R., Mullis, I.V.S. & Cotter, K.E. (2017). TIMSS 2019 mathematics framework. In I.V.S. Mullis & M.O. Martin (Eds.), *TIMSS 2019 Assessment Frameworks* (pp. 11–25). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <http://timssandpirls.bc.edu/timss2019/frameworks/>.
- Lissetz, R. & Hou, X. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology*, 13(3).
- Lovett, M.C. (2002). Problem solving. In D. Medin (Ed.), *Stevens' Handbook of Experimental Psychology: Memory and Cognitive Processes* (pp. 317–362). New York, NY: Wiley.
- Martin, M.O. & Mullis, I.V.S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

- Martin, M.O., Mullis, I.V.S, Foy, P. & O'Dwyer, L.M. (2013). Effective schools in reading, mathematics, and science at the fourth grade. In M.O. Martin & I.V.S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among Reading, Mathematics, and Science Achievement at the Fourth Grade—Implications for Early Learning* (pp.109–178). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Martin, M.O., Mullis, I.V.S. & Hooper, M. (Eds.). (2016). *Methods and Procedures in TIMSS 2015*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <http://timssandpirls.bc.edu/publications/timss/2015-methods.html>.
- 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). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <http://timssandpirls.bc.edu/timss2019/frameworks/>.
- Massachusetts Department of Elementary and Secondary Education [DESE] (2013). *2013 MCAS and MCAS-Alt technical report*. Retrieved from: <http://www.doe.mass.edu/mcas/tech/?section=techreports>.
- Massachusetts Department of Elementary and Secondary Education [DESE] (2016). *2016 MCAS and MCAS-Alt technical report*. Retrieved from: <http://www.doe.mass.edu/mcas/tech/?section=techreports>.
- Mayer, R.E. & Alexander, P.A. (Eds.). (2016). *Handbook of research on learning and instruction*. Retrieved from: <https://ebookcentral.proquest.com>.
- McDonald, R. (1982). Linear versus models in item response theory. *Applied Psychological Measurement*, 6(4), 379–396.
- McDonald, R. (1999). *Test theory: A unified treatment*. Mahwah, NJ: L. Erlbaum Associates.
- Measured Progress/ETS Collaborative (2012). *Smarter Balanced Assessment Consortium: Technology enhanced items*. Retrieved from: <https://www.measuredprogress.org/wp-content/uploads/2015/08/SBAC-Technology-Enhanced-Items-Guidelines.pdf>.
- Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.

- Messick, S. (1990). Validity of test interpretations and use. In M.C. Alkin (Ed.) *Encyclopedia of Educational Research* (6th ed.). New York, NY: MacMillan.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23.
- Mislevy, R.J., Almond, R.G. & Lukas J.F. (2003). *A brief introduction to evidence-centered design*. Retrieved from: <https://www.ets.org/Media/Research/pdf/RR-03-16.pdf>.
- Mislevy, R.J. & Riconscente, M.M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology (PADI Technical Report 9)*. Menlo Park, CA: SRI International and University of Maryland. Retrieved from: http://padi.sri.com/downloads/TR9_ECD.pdf.
- Mullis, I.V.S., Cotter, K.E., Fishbein, B.G. & Centurino, V.A.S. (2016). Developing the TIMSS 2015 achievement items. In M.O. Martin, I.V.S. Mullis & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 1.1–1.22). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <http://timss.bc.edu/publications/timss/2015-methods/chapter-1.html>.
- Mullis, I.V.S. & Martin, M.O. (Eds.). (2017). *TIMSS 2019 Assessment Frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <http://timssandpirls.bc.edu/timss2019/frameworks/>.
- Mullis, I.V.S., Martin, M.O., Cotter, K.E. & Centurino, V.A.S. (2017). *TIMSS 2019 item writing guidelines*. Retrieved from Boston College, TIMSS & PIRLS International Study Center.
- Mullis, I.V.S., Martin, M.O., Goh, S. & Cotter, K.E. (Eds.) (2016). *TIMSS 2015 Encyclopedia: Education Policy and Curriculum in Mathematics and Science*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <http://timssandpirls.bc.edu/timss2015/encyclopedia/>.
- Muthén, B. & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors*, 31, 1050–1066.
- Muthén, B. & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal likert variables. *British Journal of Mathematical and Statistical Psychology*, 38, 171–189.

- Muthén, B. & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology*, 45, 19–30.
- Muthén, L. & Muthén, B. (1998–2017). *Mplus User's Guide*. (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Muthén, B., Muthén, L. & Asparauhov, T. (2015). *Estimator choices with categorical outcomes*. Retrieved from: <https://www.statmodel.com/>.
- Muthén, B., Kaplan, D. & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52(3), 431–462.
- Muthén, B. & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, 25, 267–316.
- National Assessment of Educational Progress [NAEP] (2014a). *Technology and engineering literacy framework for the 2014 National Assessment of Educational Progress*. Retrieved from: <https://nces.ed.gov/nationsreportcard/tel/>.
- National Assessment of Educational Progress [NAEP] (2014b). *Technology and engineering literacy (TEL): Explore a scenario-based task*. Retrieved from: <https://nces.ed.gov/nationsreportcard/tel/>.
- National Assessment of Educational Progress [NAEP] (2017). *Scoring*. Retrieved from: <https://nces.ed.gov/nationsreportcard/tdw/scoring/>.
- National Council of Teachers of Mathematics [NCTM] Research Committee (2013). New assessments for new standards: The potential transformation of mathematics education and its research implications. *Journal for Research in Mathematics Education*, 44(2), 340–352.
- Nijlen, D. Van & Janssen, R. (2015). Examinee non-effort on contextualized and non-contextualized mathematics items in large-scale assessments. *Applied Measurement in Education*, 28(1), 68–84.
- Novick, L.R. & Bassok, M. (2005). Problem solving. In K.J. Holyoak & R.G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 321–349). New York, NY: Cambridge University Press.
- 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, 160–175.

- Organisation for Economic Co-operation and Development [OECD] (2017). *PISA 2015 collaborative problem solving frameworks*. Retrieved from: <https://oecd.org>.
- Padilla, J.L. & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, 26(1), 136–144.
- Parshall, C. & Becker, K. (2015). *Beyond the technology: Developing innovative items*. Pearson. Retrieved from: https://www.researchgate.net/publication/265667092_Beyond_the_Technology_Developing_Innovative_Items on December 27 2017.
- Parshall, C.G., Harmes, J.C., Davey, T. & Pashley, P.J. (2010). Innovative item types for computerized testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 215–230). New York, NY: Springer.
- Pellegrino, J.W., Chudowsky, N. & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.
- Pellegrino, J.W. & Quellmaz, E.S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education*, 43(2), 119–134.
- Poggio, J., Glasnapp, D.R., Yang, X. & Poggio, A.J. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment*, 3(6).
- Polya, G. (1957). *How to Solve It*. Garden City, NY: Doubleday.
- Quellmalz, E.S. & Pellegrino, J.W. (2009). Technology and testing. *Science*, 323(5910), 75–79.
- Reckase, M.D. (2009). *Multidimensional item response theory*. Dordrecht, Netherlands: Springer.
- Redecker, C. (2013). *The use of ICT for the assessment of key competences*. Retrieved from: <https://publications.europa.eu/>.
- Reise, S.P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667–696.
- Rijmen, F. (2009). *Efficient full information maximum likelihood estimation for multidimensional IRT models* (ETS RR-09-03). Princeton, NJ: Educational Testing Service.

- Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, 47(3), 361–372.
- Rijmen, F. (2011). Hierarchical factor item response theory models for PIRLS: capturing clustering effects at multiple levels. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments* (Vol. 4), 59–74.
- Rijmen, F., Jeon, M., von Davier, M. & Rabe-Hesketh, S. (2014). A third-order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment. *Journal of Educational and Behavioral Statistics* 39(4), 235–256.
- Robitaille, D.F., Beaton, A.E. & Plomp, T. (2000). *The Impact of TIMSS on the teaching and learning of mathematics and science*. Vancouver, Canada: Pacific Educational Press.
- Russell, M. (2016). A framework for examining the utility of technology-enhanced items. *Journal of Applied Testing Technology*, 17(1), 20–32.
- Russell, M., Goldberg, A. & O'Connor, K. (2003). Computer-based testing and validity: a look back into the future. *Assessment in Education*, 10(3), 278–293.
- Rust, K. (2014). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 117–153). Boca Raton, FL: CRC Press, Taylor & Francis Group.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 17. Richmond, VA: Psychometric Society.
- Sandene, B., Bennett, R.E., Braswell, J. & 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. Washington, DC: U.S. Government Printing Office.
- Scalise, K. (2012). *Using technology to assess hard-to-measure constructs in the common core state standards and to expand accessibility*. Paper presented at the Invitational Research Symposium on Technology Enhanced Assessment, May 7–8, 2012. Retrieved from: <https://www.ets.org/Media/Research/pdf/session1-scalise-paper-2012.pdf>.

- Scalise, K. & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing “Intermediate Constraint” questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment*, 4(6).
- Scherrer, J. (2015). Learning, teaching, and assessing the standards for mathematical practice. In C. Suurtamm & A.R. McDuffie (Eds.), *Annual perspectives in mathematics education: Assessment to enhance teaching and learning* (pp. 199–208). Reston, VA: National Council of Teachers of Mathematics.
- Schoenfeld, A.H. (1985). *Mathematical problem solving*. Orlando, FL: Academic Press Inc.
- Schwarz, G. (1978). Estimating the dimension of a model. *AnnStat*, 6, 461–464.
- Shu, Z., Bergner, Y. Zhu, M., Hao, J. & von Davier A. (2017). An item response theory analysis of problem- solving processes in scenario-based tasks. *Psychological Test and Assessment Modeling*, 59(1), 109–131.
- Sireci, S.G. & Zenisky, A. (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. 329–376). Mahwah, NJ: Lawrence Erlbaum Associates.
- Smarter Balanced Assessment Consortium (2016). *Smarter Balanced Assessment Consortium: 2014-15 Technical Report*. Retrieved from: http://www.smarterbalanced.org/wp-content/uploads/2015/08/2013-14_Technical_Report.pdf.
- Strain-Seymour, E., Way, W.D. & Dolan, R.P. (2009). *Strategies and processes for developing innovative items in large-scale assessments*. Iowa City, IA: Pearson Education.
- Stein, M.K. & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation*, 2(1), 50–80.
- Stigler, J.W. & Hiebert, J. (2004). Improving mathematics teaching. *Educational Leadership*, 61(5), 12–17.
- Sugrue, B. (1995). A theory-based framework for assessing domain-specific problem solving ability. *Educational Measurement Issues and Practice*, 14(3), 29–35.

- Suurtamm, C., Thompson, D.R., Kim, R.Y., Moreno, L.D., Sayac, N., Schukajlow, S., Silver, E. ...Vos, P. (2016). Assessment in mathematics education: Large-scale assessment and classroom assessment. In K. Gabriele (Ed.), *ICME-13 topical surveys* (pp. 1–38). Springer International Publishing: Springer.
- Swan, M. & Burkhardt, H. (2012). A designer speaks: Designing assessment of performance in mathematics. *Educational Designer: Journal of the International Society for Design and Development in Education*, 2(5), 1–41.
- Threlfall, J., Pool, P., Homer, M. & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics*, 66(3), 335–338.
- TIMSS & PIRLS International Study Center (2017). *About TIMSS 2019*. Retrieved from: <https://timssandpirls.bc.edu/timss2019/>.
- Toland, M.D., Sulis, I., Giambona, F., Porcu, M. & Campbell, J.M. (2017). Introduction to bifactor polytomous item response theory analysis. *Journal of School Psychology*, 60, 41–63.
- Walcott, C.Y., Hudson, R., Mohr, D. & Essex, N.K. (2015). What NAEP tells us about the relationship between classroom assessment practices and student achievement in mathematics. In C. Suurtamm & A. Roth McDuffie (Eds.), *Annual perspectives in mathematics education: Assessment to enhance teaching and learning* (pp. 179–190). Reston, VA: National Council of Teachers of Mathematics.
- Wang, J. & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. John Wiley & Sons, Incorporated. Retrieved from ProQuest Ebook Central.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. New York, NY: Psychology Press.
- Wise, S.L., Pastor, D.A. & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice, *Applied Measurement in Education*, 22(2), 185–205.
- Wu, J. & Kwok, O. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches, *Structural Equation Modeling: A Multidisciplinary Journal*, 19(1), 16–35.
- Yamamoto, K., He, Q., Shin, H.J. & von Davier, M. (2017). *Developing a machine-supported coding system for constructed-response items in PISA*. ETS Research Report Series, 2017: 1–15.

Appendix

Appendix A: eTIMSS Mathematics PSI Development

Milestones

Exhibit A.1: eTIMSS Mathematics PSI Development Milestones, January 2015–September 2018

Date		Group and Activity
January	2015	<p>TIMSS & PIRLS International Study Center and IEA Hamburg began preparing for the transition to eTIMSS:</p> <ul style="list-style-type: none"> • TIMSS & PIRLS International Study Center began planning to convert trend items to digital format and develop new items, including the PSIs, for tablet-based delivery. • IEA Hamburg began designing the eTIMSS Infrastructure.
March	2015	<p>Initial PSI task development began under the assumption that the platform would be tablet and stylus to replicate paper-and-pencil. The countries were responsible for providing the devices.</p> <ul style="list-style-type: none"> • TIMSS & PIRLS International Study Center began work with members of the Science and Mathematics Item Writing Committee (SMIRC), other external expert consultants, and IEA Hamburg to design and operationalize prototype PSIs based on the <i>TIMSS 2015 Framework</i>. • Initial development goals were established, including the characteristics of a successful PSI. • By August 2015, one fourth grade mathematics PSI was fully operationalized.
August	2015	<p>Consultants and staff at the TIMSS & PIRLS International Study Center began drafting additional PSIs for both the fourth and eighth grade. (Boston, USA)</p>
August	2015	<p>American Institute for Research (AIR) conducted cognitive laboratories for two PSIs (one fourth grade mathematics and one eighth grade science) and a sample of TIMSS trend items converted to digital format.</p>
October	2015	<p>Consultants and staff at the TIMSS & PIRLS International Study Center continued revising the draft PSIs and drafted new tasks. Staff at the TIMSS & PIRLS International Study Center updated the group on recent advances in eTIMSS plans, including the updated computer or tablet design and changes to the user interface. The group also revisited the characteristics of a successful PSI. (Boston, USA)</p>

Date		Group and Activity
June	2016	TIMSS & PIRLS International Study Center presented an informational video introducing the features of the eTIMSS assessments and debuting the PSIs to NRCs at the 8 th TIMSS 2015 NRC Meeting. (Quebec, Canada)
September	2016	SMIRC reviewed a draft of the <i>TIMSS 2019 Mathematics Framework</i> and provided feedback to staff at the TIMSS & PIRLS International Study Center. The staff then met with SMIRC consultants to incorporate the SMIRC's comments. In updating the framework, the group paid special attention to the novel affordances of eTIMSS for assessing traditionally difficult to measure areas of mathematics. (Boston, USA)
October	2016	Australia, Canada, and Singapore administered draft PSIs in the eTIMSS prePilot. The mathematics prePilot instruments included four PSIs at the fourth grade and three at the eighth grade.
November	2016	Consultants and staff at the TIMSS & PIRLS International Study Center reviewed the results of the eTIMSS prePilot and revised the tasks based on these results. The group also drafted one additional PSI for each grade, fulfilling the development requirements for the eTIMSS 2019 Field Test. (Boston, USA)
December	2016	To prepare for the Field Test, the TIMSS & PIRLS International Study Center and IEA Hamburg ramped up programming efforts, including both front- and back-end development. This work continued for over a year and included extensive quality assurance and some additional revisions to the PSIs.
February	2017	NRCs reviewed the draft <i>TIMSS 2019 Mathematics Framework</i> at the 1 st TIMSS 2019 NRC Meeting (Hamburg, Germany). Following the meeting, the NRCs completed an online survey through which they provided feedback about whether each mathematics topic area should be kept as is, modified, or deleted.
April	2017	SMIRC reviewed both the draft <i>TIMSS 2019 Mathematics Framework</i> and the fourth and eighth grade PSIs at the 1 st SMIRC meeting. (Amsterdam, The Netherlands)
May	2017	The eTIMSS Pilot / Item Equivalence Study, designed to investigate mode effects for the TIMSS trend items, was conducted in 24 countries at the fourth grade and 13 countries at the eighth grade. This study did not include any PSIs, but gave valuable information about the robustness of the eAssessment Systems and countries readiness to conduct a digital assessment.

Date	Group and Activity
September 2017	Consultants and staff at the TIMSS & PIRLS International Study Center reviewed the updated PSIs and refined the scoring guides with special attention to machine scoring. The group also began to discuss what event data (e.g., use of tools, going back to previous screens) would be of interest for future analyses. (Boston, USA)
November 2017	NRCs reviewed the Field Test PSIs at the 3 rd NRC Meeting. (Melbourne, Australia)
December 2017	TIMSS & PIRLS International Study Center and IEA Hamburg finalized all eTIMSS 2019 Field Test instruments and released the international instruments to countries for translation.
January 2018	TIMSS & PIRLS International Study Center and IEA Hamburg collaborated to establish specifications for data capture and scoring. The specifications were finalized in March 2018.
January 2018	Consultants and staff at the TIMSS & PIRLS International Study Center reviewed the field test scoring guides and prepared scorer training materials. (Boston, USA)
March 2018	Countries conducted the eTIMSS 2019 Field Test in March – May 2018.
March 2018	NRCs received scoring training for constructed response items at the 4 th NRC meeting. (Madrid, Spain)
April 2018	The author observed four eTIMSS 2019 Field Test sessions in the United States and prepared a report documenting these observations for the TIMSS & PIRLS International Study Center.
May 2018	Countries submitted eTIMSS 2019 Field Test achievement data for analysis and review.
May 2018	NRCs provided feedback on the field test PSIs to the TIMSS & PIRLS International Study Center. Based on the NRC's evaluations, the TIMSS & PIRLS International Study Center selected the PSIs to move forward to eTIMSS 2019 Data Collection and began editing the tasks based on NRC feedback.
June 2018	IEA Hamburg completed data processing and the TIMSS & PIRLS International Study Center completed scoring of machine-scored items.
June 2018	TIMSS & PIRLS International Study Center reviewed field test achievement item almanacs and selected the items for data collection.

Date		Group and Activity
July	2018	SMIRC reviewed the proposed items for data collection in conjunction with the field test results at the 3 rd SMIRC meeting. (Tromsø, Norway)
August	2018	NRCs reviewed and approved item blocks for TIMSS 2019 Data Collection at the 5 th NRC meeting. (Stockholm, Sweden)
September	2018	TIMSS & PIRLS International Study Center and IEA Hamburg finalized all eTIMSS 2019 Data Collection instruments and released the international instruments to countries for translation. Data collection began in the Southern Hemisphere.

Appendix B: Selected Survey Activities Questionnaire Items

Selected items from the eTIMSS 2019 Field Test Survey Activities Questionnaire that were considered in this dissertation. The questionnaire was made available to NRCs in April 2018. NRCs from 22 eTIMSS countries responded to the questionnaire.

Preparing Instruments

1. Did you experience any problems receiving the eTIMSS Player(s) from IEA Hamburg and preparing the eTIMSS USB sticks/tablets? If yes, please specify.
2. In your opinion, was there any information that was missing, or sections that could be shortened or omitted from the international version of the “Preparing Computers and/or Tablets for eTIMSS” instructions? If yes, please specify.
3. Did you experience any software specific problems when using the eTIMSS System Check Program to test computers/tablets for eTIMSS compatibility? If yes, please specify.

Conducting Testing Sessions

4. In your opinion, was there any information that was missing, or sections that could be shortened or omitted from the international versions of the School Coordinator Manual or Test Administrator Manual?
5. Did you require/suggest/provide an additional person to help the Test Administrators during the eTIMSS testing sessions? If yes, please describe the situation and whether you found this help necessary. Would you consider this for the main data collection?
6. Please briefly summarize any problems or special circumstances of the test administration that were documented in the Test Administration Forms by the Test Administrators.
7. Did you experience any software specific problems when using the eTIMSS Player(s)? If yes, please specify.
8. Did you experience any problems with the eTIMSS Online Data Monitor? If yes, please specify.
9. Please briefly summarize the activities of your national quality control program and any problems encountered by the monitors.

Scoring Student Responses

10. Did you encounter any problems with the scoring materials provided by the TIMSS & PIRLS International Study Center? Did you translate any of these materials, or create additional national scoring training materials?
11. Please specify any problems you encountered while training your scorers and/or during the scoring process.
12. Did you experience any problems with the eTIMSS Scoring System? If yes, please specify.

Appendix C: eTIMSS Student Questionnaire Results

Exhibit C.1: International Summary Statistics for Students Like Taking the Test on a Computer or Tablet

Grade	Valid Cases	Percentages				
		Liked it a lot	Liked it a little	Didn't like it very much	Didn't like it at all	Missing
Grade 4	42,318	66.8	27.3	4.0	1.9	0.4
Grade 8	30,406	43.5	38.1	11.4	7.1	0.4

Exhibit C.2: International Summary Statistics for Students Experiencing Difficulties Taking eTIMSS

	Grade 4 Percentages				Grade 8 Percentages			
	Valid Cases	Yes	No	Missing	Valid Cases	Yes	No	Missing
It was hard to type	41,262	18.3	81.7	0.3	29,852	18.5	81.5	0.2
I had trouble using the number pad	40,914	10.3	89.7	0.3	29,810	21.5	78.5	0.2
Objects were hard to drag	40,863	11.5	88.5	0.2	29,634	12.9	87.1	0.2
There was no good place to work out my answers	40,684	14.8	85.2	0.3	29,659	19.6	80.4	0.2
The computer or tablet was slow	40,723	12.6	87.4	0.2	29,663	15.7	84.3	0.1
I had to start my test over because of a computer or tablet problem	40,260	5.3	94.7	0.4	29,459	5.9	94.1	0.4

Appendix D: Average Time per Screen for the Mathematics PSIs in the eTIMSS Field Test

Exhibit D.1: Average Time per Screen in the eTIMSS Field Test – Grade 4 PSI Blocks

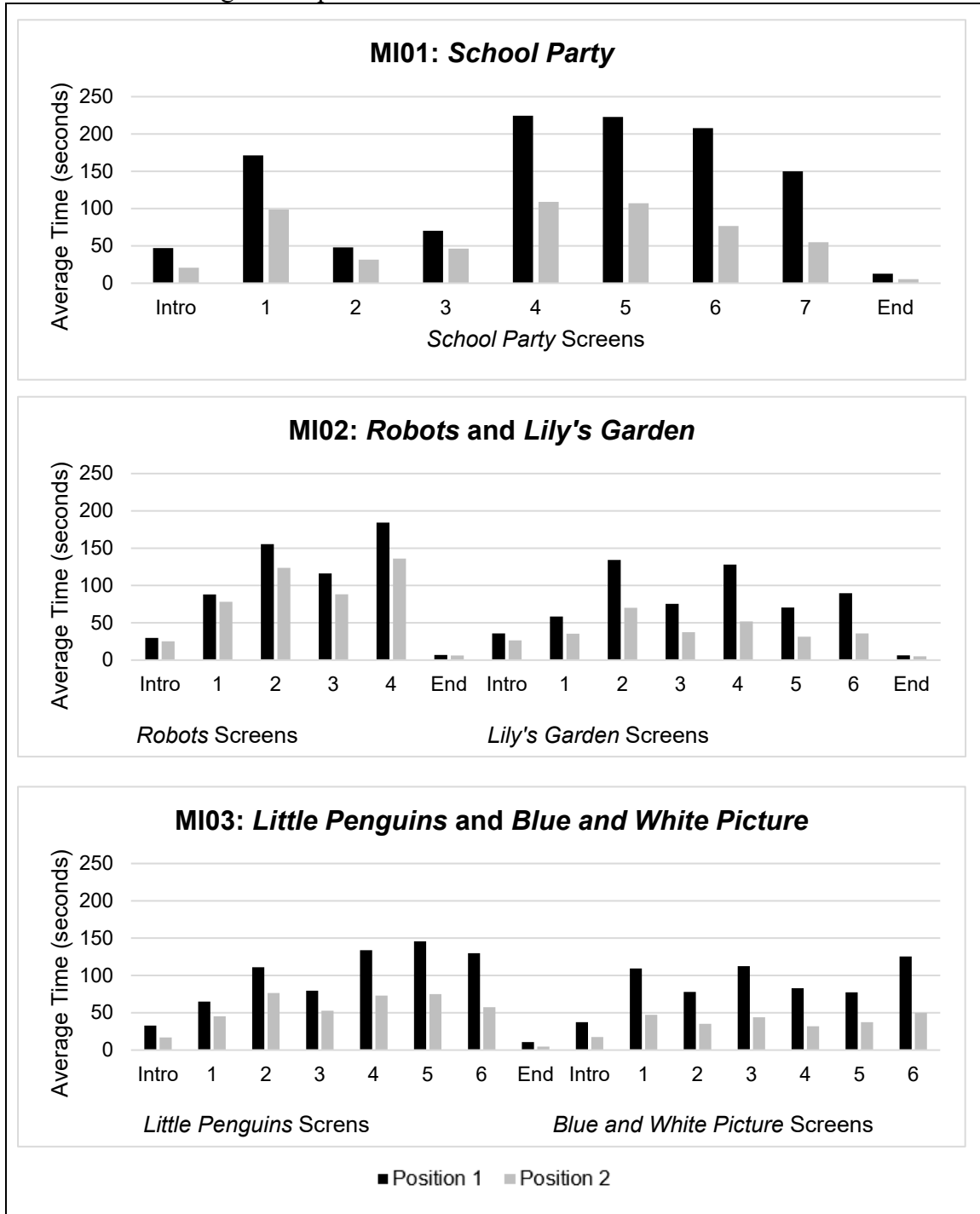


Exhibit D.2: Average Time per Screen in the eTIMSS Field Test – Grade 8 PSI Blocks

