Preserving 20 Years of TIMSS Trend Measurements: Early Stages in the Transition to the eTIMSS Assessment

Author: Bethany Fishbein

Boston College
Lynch School of Education

Department of
Measurement, Evaluation, Statistics, and Assessment

PRESERVING 20 YEARS OF TIMSS TREND MEASUREMENTS: EARLY STAGES IN THE TRANSITION TO THE eTIMSS ASSESSMENT

Dissertation
by

BETHANY FISHBEIN

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

May 2018

Preserving 20 Years of TIMSS Trend Measurements: Early Stages in the Transition to the eTIMSS Assessment

by

Bethany Fishbein

Dissertation Director: Ina V.S. Mullis

Abstract

This dissertation describes the foundation for maintaining TIMSS' 20 year trend measurements with the introduction of a new computer- and tablet-based mode of assessment delivery—eTIMSS. Because of the potential for mode effects on the psychometric behavior of the trend items that TIMSS relies on to maintain comparable scores between subsequent assessment cycles, development efforts for TIMSS 2019 began over three years in advance. This dissertation documents the development of eTIMSS over this period and features the methodology and results of the eTIMSS Pilot / Item Equivalence Study. The study was conducted in 25 countries and employed a within-subjects, counterbalanced design to determine the effect of the mode of administration on the trend items. Further analysis examined score-level mode effects in relation to students' socioeconomic status, gender, and self-efficacy for using digital devices. Strategies are discussed for mitigating threats of construct irrelevant variance on students' eTIMSS performance.

The analysis by student subgroups, similar item discriminations, high cross-mode correlations, and equivalent rankings of country means provide support for the equivalence of the mathematics and science constructs between paperTIMSS and

eTIMSS. However, the results revealed an overall mode effect on the TIMSS trend items, where items were more difficult for students in digital formats compared to paper. The effect was larger in mathematics than science. An approach is needed to account for the mode effects in maintaining trend measurements from previous cycles to TIMSS 2019. Each eTIMSS 2019 trend country will administer the paper trend booklets to an additional nationally representative bridge sample of students, and a common population equating approach will ensure the link between paperTIMSS and eTIMSS scores.

Acknowledgments

I must thank Dr. Ina V.S. Mullis, my dissertation adviser, and Dr. Michael O. Martin, reader for helping make this dissertation possible. I am incredibly grateful for the opportunity and guidance from the fearless leaders of the TIMSS & PIRLS International Study Center.

Thank you to Dr. Zhushan Li ("Mandy"), for serving as reader on my dissertation committee and for giving a unique perspective on this project. I am extremely grateful for your encouragement and advice in communicating technical detail. I would additionally like to thank Measurement, Evaluation, Statistics, and Assessment (MESA) faculty Dr. Larry Ludlow and Dr. Michael Russell, who provided support and guidance during the early writing stages.

From the TIMSS & PIRLS International Study Center, I would like to thank Pierre Foy for guiding the analysis for the eTIMSS Pilot / Item Equivalence Study and helping ensure the accuracy of my technical writing. Thank you for your support over the past year. Thank you to Liqun Yin for your work on the item analysis and to Yenileis Pardini for helping me write about designing the eTIMSS user interface. Thank you also to Kerry Cotter, Martin Hooper, and Caroline Prendergast for talking through ideas and problems encountered in my research.

To my best friend and colleague, Erin Wry, whom would not be in my life without our shared passion for large-scale international assessment—thank you for pushing me to be my best every single day. Thank you for reading over drafts and for helping with the item analysis and survey operations for the eTIMSS Pilot / Item Equivalence Study.

Thank you to my family—Mom (Carol), Dad (Danny), and Mike Fishbein, for your love and support and for keeping up with the ups and downs of being a doctoral student. Thank you to the Foley, Launse, Shapiro, and Centner families for your encouragement. Thank you to my friends around the country who have kept in touch these last several years. In particular, I am grateful to Kelly Cupo, Sam Ricker, Jess Wala, Kimia Mavon, and Emily Bourque for coming over on week nights and keeping me entertained. Ally McAndrews Washo—thank you for always answering your phone and reminding me I am not alone.

Finally, I'd like to thank the many individuals and organizations involved in conducting the eTIMSS Pilot / Item Equivalence Study, including staff from the TIMSS & PIRLS International Study Center, IEA Hamburg, and Educational Testing Service as well as the National Research Coordinators, schools, and students from each of the participating countries.

Table of Contents

Chapter 1: Introduction 1
Innovations for TIMSS 2019 2
Challenges in Transitioning to a Computer-based Assessment 3
Measurement Challenges 3
Accommodating for Country Diversity 5
Learning from Other Assessments 5
The Path to eTIMSS 9
Description of Dissertation 10
Chapter 2: eTIMSS Development History - October 2014–December 2017 15
Overview: Major Development Milestones 15
A Note about Problem Solving and Inquiry Tasks (PSIs) 18
Planning the eAssessment System for TIMSS 19
Converting Paper Trend Items to a Tablet-and-Stylus Format 20
Tablet-and-Stylus Trend Item Classifications 21
Designing the User Interface for eTIMSS 21
Layout 23
Navigation 24
Assessment Tools and Response Actions 24
AIR Cognitive Labs - August–September 2015 25
Feasibility Constraints: Accommodating Country Diversity 26
eTIMSS prePilot - September–October 2016 28
eTIMSS Pilot / Item Equivalence Study - January–June 2017 33
Overview of the eTIMSS Pilot / Item Equivalence Study 33
Preparing Paper Booklets 35
Tracking Class/Student Participation 36
Preparing Computers for eTIMSS 38
Adding Trend Translations to the eTIMSS Translation System 39
Test Administrator Manuals for paperTIMSS and eTIMSS 41
Scoring the Constructed Response Items 43

Entering and Submitting the Tracking and paperTIMSS Data 45
Feedback from Country Participants 46
Preparing for the Field Test - June 2017–April 2018 47
Updating the eTIMSS Online Translation System 48
Improving the eTIMSS Player 49
Writing Manuals 50
Developing Digitally Enhanced Items 52
Chapter 3: Methodology and Results of the eTIMSS Pilot / Item Equivalence Study 55
Overview 55
Sample Design 56
Instruments 56
Achievement Booklets / Block Combinations 56
Student Questionnaire 60
Research Design 60
Phase 1: A Priori Analysis of Item Equivalence 62
A Priori Analysis Procedure 62
A Priori Analysis Results 66
Phase 2: Item Analysis 67
Item Analysis Procedure 68
Reviewing the Item Statistics 68
Refining the A Priori Item Hypotheses 72
Producing International Average Item Statistics 75
Item Analysis Results 76
Results by Digital Item Type 81
Item Analysis Summary 85
Phase 3: Scale Score Analysis 87
Methodological Overview 87
Scale Score Analysis Procedure 88
Calibrating the Items 88
Estimating Scale Scores 90
Producing Statistics to Examine Score Comparability 91

Scale Score Analysis Results 93
Scale Score Analysis Summary 97
Discussion of the Results of the eTIMSS Pilot / Item Equivalence Study 98
Measuring Trends in TIMSS 99
Linking paperTIMSS and eTIMSS Scores for TIMSS 2019 102
Conclusion 104
Chapter 4: Analysis of the eTIMSS Item Equivalence Database 105
Predictors of Testing Mode Effects: A Review of the Literature 106
Measures of Socioeconomic Status 107
Gender 107
Computer and Tablet Experience 108
Research Summary 110
Analysis Methodology 111
Description of the eTIMSS Item Equivalence Database 111
Variables of Interest 113
Gender 113
Socioeconomic Status 113
Digital Self-Efficacy 115
Analysis Procedures 117
Phase 1 Analysis Procedures 117
Phase 2 Analysis Procedures 119
Results 121
Phase 1 Results 121
Socioeconomic Status 122
Gender 127
Digital Self-Efficacy 129
Phase 2 Results 132
Discussion of the Results 135
Limitations and Suggestions for Further Research 138
Chapter 5: Discussion 139
Overview of Dissertation 139
eTIMSS Pilot / Item Equivalence Study 140

Analysis of the eTIMSS Item Equivalence Database 140
Major Findings 142
Preserving TIMSS Trend Measurements: Next Steps 147
Linking paperTIMSS and eTIMSS with Common Population Equating 147
Bridge Samples 148
Improving the Trend Items 149
Re-assessing Trend Item Equivalence 149
Additional Recommendations for TIMSS 149
eTIMSS Direction Module 150
Collecting User Experience Data 150
Measuring Computer Experience and Attitudes 151
References 152
Appendix A: Constructing the Digital Self-Efficacy Scale 163
Selecting Scale Items 163
Evaluating Unidimensionality 166
Calibrating the Items 168
Assessing Item Dependence 170
Producing Scale Scores 172
Validating the Scale 173
Creating Content-Referenced Regions 174

List of Exhibits

Exhibit 2.1: eTIMSS Development Milestones, October 2014–December 2017 16

Exhibit 2.2: Example Item Screen from the eTIMSS Pilot / Item Equivalence Study 31

Exhibit 2.3: Example Student Tracking Form for the eTIMSS Pilot / Item Equivalence Study 37

Exhibit 2.4: eTIMSS 2019 Pilot / Item Equivalence Study Minimum Device Requirements* 38

Exhibit 3.1: Countries Participating in the eTIMSS Pilot / Item Equivalence Study 56

Exhibit 3.2: Distribution of Items in the eTIMSS Pilot / Item Equivalence Study by Digital Item Type 59

Exhibit 3.3: eTIMSS Pilot / Item Equivalence Study Booklet/Block Combination Design—Fourth and Eighth Grades 61

Exhibit 3.4: eTIMSS Pilot / Item Equivalence Study Booklet/Block Combination Rotation Scheme—Fourth and Eighth Grades 62

Exhibit 3.5: A Priori Trend Item Classifications for the eTIMSS Pilot / Item Equivalence Study 65

Exhibit 3.6: Distribution of Items in the eTIMSS Pilot / Item Equivalence Study by A Priori Classification 66

Exhibit 3.7: Example Item Statistics for a Multiple-Choice Item—paperTIMSS 69

Exhibit 3.8: Example Item Statistics for a Constructed Response Item—eTIMSS 69

Exhibit 3.9: Example Difference Statistics for a Compound Multiple-Choice Item 71

Exhibit 3.10: Item Plots of International Average Percent Correct Statistics for *Expected Non-Invariant* Items—paperTIMSS vs. eTIMSS 75

Exhibit 3.11: eTIMSS Item Equivalence Database—Student and Item Sample Sizes* 76

Exhibit 3.12: Item Plots of International Average Percent Correct Statistics for *Expected Invariant* Items—paperTIMSS vs. eTIMSS 77

Exhibit 3.13: International Average Percent Correct Statistics (Item Difficulty) 79

Exhibit 3.14: International Average Point-Biserial Correlations (Item Discrimination) 80

Exhibit 3.15: International Average Percent Omitted Statistics 80

Exhibit 3.16: International Average Percent Not Reached Statistics 81
Exhibit 3.17: International Average Percent Correct Statistics (Item Difficulty) by Digital Item Type—Fourth Grade, Mathematics 82
Exhibit 3.18: International Average Percent Correct Statistics (Item Difficulty) by Digital Item Type—Fourth Grade, Science 83
Exhibit 3.19: International Average Percent Correct Statistics (Item Difficulty) by Digital Item Type—Eighth Grade, Mathematics 84
Exhibit 3.20: International Average Percent Correct Statistics (Item Difficulty) by Digital Item Type—Eighth Grade, Science 85
Exhibit 3.21: International Average Scale Scores and Standard Deviations 94
Exhibit 3.22: International Average Scale Scores, Standard Errors, and Mode Effect Sizes 94
Exhibit 3.23: Country Distribution of Average Scale Score Differences between paperTIMSS and eTIMSS 96
Exhibit 3.24: International Average Cross-Mode Correlation Coefficients (Adjusted for Reliability) 97
Exhibit 3.25: Concurrent Calibration Model Used for TIMSS Trend Measurements 99
Exhibit 3.26: Concurrent Calibration Model for TIMSS 2019 103
Exhibit 4.1: eTIMSS Item Equivalence Database—Percentage of Students by Gender (Unweighted) 112
Exhibit 4.2: eTIMSS Item Equivalence Database—Percentage of Students by eTIMSS Device (Unweighted) 113
Exhibit 4.3: eTIMSS Pilot / Item Equivalence Study eTIMSS Questionnaire Items Measuring Socioeconomic Status 114
Exhibit 4.4: eTIMSS Pilot / Item Equivalence Study eTIMSS Questionnaire Items Measuring Digital Self-Efficacy 116
Exhibit 4.5: International Average Scale Scores by Books in the Home—Fourth Grade (21 countries) 122
Exhibit 4.6: International Average Scale Scores by Books in the Home—Eighth Grade (11 countries) 124
Exhibit 4.7: Average Mathematics and Science Mode Effects by Socioeconomic Status 125
Exhibit 4.8: International Average Scale Scores by Parents' Highest Level of Education—Eighth Grade (11 countries) 126

Exhibit 4.9: International Average Scale Scores by Gender—Fourth Grade (21 countries) 127
Exhibit 4.10: International Average Scale Scores by Gender—Eighth Grade (11 countries) 128
Exhibit 4.11: Average Mathematics and Science Mode Effects by Gender 129
Exhibit 4.12: International Average Scale Scores by Level of Digital Self-Efficacy—Fourth Grade (21 countries) 130
Exhibit 4.13: International Average Scale Scores by Level of Digital Self-Efficacy—Eighth Grade (11 countries) 131
Exhibit 4.14: Average Mathematics and Science Mode Effects by Digital Self-Efficacy 131
Exhibit 4.15: International Average Regression Coefficients—Fourth Grade (21 countries) 133
Exhibit 4.16: International Average Regression Coefficients—Eighth Grade (11 countries) 134
Exhibit 4.17: International Average Regression Coefficients of Socioeconomic Status on Mode Effects in Science—Fourth and Eighth Grades 135
Exhibit 5.1: Summary of the Results of the eTIMSS Pilot / Item Equivalence Study 144
Exhibit 5.2: Distribution of Mathematics and Science Mode Effects Across Countries 145
Exhibit A.1: International Summary Statistics for Items Measuring Digital Self-Efficacy—Fourth Grade 164
Exhibit A.2: International Summary Statistics for Items Measuring Digital Self-Efficacy—Eighth Grade 165
Exhibit A.3: Principal Components Analysis of the Digital Self-Efficacy Scale—Fourth and Eighth Grades 167
Exhibit A.4: International Item Parameters for the Digital Self-Efficacy Scale—Fourth Grade 169
Exhibit A.5: International Item Parameters for the Digital Self-Efficacy Scale—Eighth Grade 169
Exhibit A.6: Autocorrelation Statistics for the Digital Self-Efficacy Scale—Fourth Grade 171
Exhibit A.7: Autocorrelation Statistics for the Digital Self-Efficacy Scale—Eighth Grade 171

Exhibit A.8: Scale Transformation Constants for the Digital Self-Efficacy Scale—Fourth and Eighth Grades 173

Exhibit A.9: International Average Correlation between Digital Self-Efficacy and Achievement 173

Exhibit A.10: Equivalence Table of Raw and Transformed Scale Scores for the Digital Self-Efficacy Scale—Fourth and Eighth Grades 175

Exhibit A.11: Percentages of Students by Level of Digital Self-Efficacy—Fourth and Eighth Grades 176

Chapter 1: Introduction

TIMSS (the Trends in International Mathematics and Science Study) is an international comparative study of student achievement in mathematics and science around the world. Conducted on a four-year assessment cycle since 1995, TIMSS has assessed student achievement at the fourth and eighth grades six times—in 1995, 1999, 2003, 2007, 2011, and 2015—and has accumulated 20 years of trend measurements. Now for the 2019 assessment cycle, TIMSS is transitioning to an innovative, computer-based "eAssessment system." Not all of the approximately 65 TIMSS countries have the infrastructure to switch to a completely computerized system, so half are transitioning in 2019 and the other half will make the change in 2023. The transition to the eAssessment system is an enormous undertaking for TIMSS. Developing the eAssessment system began with extensive planning to build a multi-component software and application system to accommodate the many processes involved in conducting a large-scale international assessment. Early stages of development also included small pre-tests as well as an item equivalence study, where the same students took the TIMSS trend items in both versions—paperTIMSS and eTIMSS. The results helped TIMSS determine plans for trend measurement for TIMSS 2019. This dissertation documents the findings of these first steps on the path to eTIMSS.

TIMSS also collects extensive data about the contexts for learning, including school climates, instructional resources, teacher practices, and student characteristics, attitudes, and home supports for learning. With this information, TIMSS has the goal of providing countries with evidence about factors that can contribute to improvements in

student achievement, with an emphasis on measuring change in education systems to inform policy.

TIMSS is directed by the TIMSS & PIRLS International Study Center at Boston College and is the flagship international comparative study conducted under the auspices of IEA (the International Association for the Evaluation of Educational Achievement). IEA has offices in Amsterdam and Hamburg. The Hamburg location houses a large research and data processing center, where the eAssessment system for TIMSS is being developed in collaboration with the TIMSS & PIRLS International Study Center.

Innovations for TIMSS 2019

Because of its ambitious nature, TIMSS began work on its eAssessment system more than three years ago. The system encompasses the capabilities for creating achievement instruments, delivering the assessment to the countries, conducting translation and translation verification, uploading data, and scoring student responses. Students from the participating eTIMSS countries will take the TIMSS 2019 assessment on personal computers (PCs) or tablets and their item responses will be uploaded directly to IEA servers. Once a labor-intensive process, the process of distributing and scoring constructed response items is much more efficient, with machine-scoring capabilities as well as web-based scoring system developed by staff at IEA Hamburg.

The shift from the traditional paper-and-pencil administration used in previous cycles to a fully computer-based testing system will ultimately be beneficial for TIMSS, providing enhanced measurement capabilities and extended coverage of the TIMSS assessment frameworks in mathematics and science. TIMSS 2019 will include extended

Problem Solving and Inquiry Tasks (PSIs), and students will have digital tools available through the eTIMSS interface, including a number pad, ruler, and calculator. Students will see a larger variety of response modes to answer digitally enhanced items, including drag and drop, sorting, and drop-down menu input types.

Challenges in Transitioning to a Computer-based Assessment

TIMSS will continue its 20 year trend measurements in 2019 while transitioning to a digital environment. The TIMSS approach to measuring trends includes retaining a substantial portion of the items (approximately 60 percent) from previous assessment cycles to be administered in the next cycle. The items from TIMSS 2015 that will be re-administered in TIMSS 2019 are called "trend items." Trend items enable the linking of TIMSS item response theory (IRT) achievement scales from cycle to cycle so that changes in student achievement can be accurately measured (Foy & Lin, 2016). Using concurrent calibration, TIMSS scales the newly collected data from each cycle with the data from the previous assessment cycle to produce IRT item parameters on a common scale. After producing student proficiency scores, linear transformations are applied to place results from this scale on the same scale as the results of the previous assessment.

Considering the many changes required in TIMSS as well as the project's complex international context, designing the procedures to transition to eTIMSS and maintain 20 years of trends involves careful attention to all aspects of the assessment.

Measurement Challenges

To accurately measure changes in student achievement from assessment to assessment, TIMSS' student achievement estimates need to be based on a large number

of items that are identical between assessment cycles to make the scores equivalent. Thus, its primary goal of trend measurement requires maintaining as much as possible the equivalence of the trend items across paper and digital delivery modes for the countries transitioning to eTIMSS.

Prior to TIMSS 2019, countries re-printed the trend items from the previous assessment cycle to administer again, maintaining the appearance and translations of the items. However, changing from paper-and-pencil to the new computer- and tablet-based administration could have substantial and unpredictable effects on the psychometric behavior of the trend items and student achievement scores (Mazzeo & von Davier, 2014). The IRT-based methodology TIMSS uses to estimate student achievement assumes that the psychometric characteristics of items and how they function are the same ("invariant") in different contexts (Lord, 1980; Mislevy, 1991). However, previous studies have found substantial differences in student assessment performance between paper and digital modes, called "mode effects" (APA, 1986; Bennett et al., 2008; Jerrim, 2016). For example, items may be easier or more difficult for students in a digital environment than on paper. These performance differences could vary systematically according to students' characteristics such as gender and their familiarity and confidence with using PCs and tablets (Bennett et al., 2008; Cooper, 2006; Gallagher, Bridgeman, & Cahalan, 2002; Horkay, Bennett, Allen, Kaplan, & Yan, 2006; Jerrim, 2016; Zhang, Xie, Park, Kim, Broer, & Bohrnstedt, 2016).

The potential impact of the digital mode of administration on student performance as reflected by the psychometric behavior of TIMSS trend items is a concern, because it may affect the comparability of TIMSS achievement scores between paper and digital

delivery modes (APA, 1986; Russell, Goldberg, & O'Connor, 2003). Therefore, early stages in developing eTIMSS for 2019 involved identifying and trying to mitigate all possible sources of mode effects to preserve trend measurements.

Accommodating for Country Diversity

TIMSS encouraged as many countries as possible to transition to eTIMSS for the 2019 cycle, and developed the eTIMSS assessment to be compatible with a variety of digital devices, so that countries can use existing digital devices whenever possible. Just recently, eTIMSS needed to accommodate the use of Google Chromebooks, and new technologies are continuously emerging. TIMSS is fully aware that accommodating for a variety of PC and tablet devices introduced the potential for further variation in student performance between modes related to device effects (Davis, Kong, McBride, & Morrison, 2017; DePascale, Dadey, & Lyons, 2016; Strain-Seymour, Craft, & Davis, 2013; Way, Davis, Keng, & Strain-Seymour, 2016). This led to considerable efforts to keep TIMSS trend items and testing procedures standard across devices.

Learning from Other Assessments

The experiences of other assessments have helped staff at the TIMSS & PIRLS International Study Center identify and mitigate several possible sources of mode effects on eTIMSS student performance. Of particular concern is the potential for computers or tablets to inhibit students' abilities to provide evidence of the constructs that the trend items mean to measure. These difficulties experienced by students could be due to their unfamiliarity with using digital devices or with the types of actions required to navigate within the eTIMSS assessment and respond to items (Duque, 2016; Johnson & Green,

2006; Winter, 2010). For example, if a student is unfamiliar with using computers or tablets—particularly in assessment contexts—more time may be spent learning the assessment interface or typing responses, rather than engaging with test content (Davis et al., 2017; Russell, 1999; Pisacreta, 2013). Conversely, students more familiar with computers may capitalize upon the technology for fluency and efficiency (Clariana & Wallace, 2002).

Research findings about mode effects have led researchers and assessment programs like TIMSS to work to minimize the potential impact of mode effects on student performance, typically estimated by IRT item parameters and achievement scores. In particular, the experiences of NAEP (the National Assessment of Educational Progress) and PISA (the Programme for International Student Assessment) provided insight into the difficulties involved in transitioning to a computer-based testing platform while trying to maintain accurate trend measurements (Bennett et al., 2008; Mazzeo & von Davier, 2008; OECD, 2015; 2016; 2017; Sandene, Bennett, Braswell, & Oranje, 2005).

PISA 2015 results showed drastic drops in student performance on computer-based delivery from that on paper in some countries, leading to concerns worldwide about the validity of PISA's trend measurements (Grzanna, 2017; Robitzsch et al., 2016; Ward, 2017). Prior to data collection, PISA strategically developed a research design and scaling procedures to maintain linkages across test forms, regardless of delivery mode (Mazzeo & von Davier, 2008; OECD 2016; 2017). However, the psychometric procedures used to link paper and digital scores proved inadequate in some circumstances (Jerrim, 2018). The differences in performance between paper and digital test forms

varied in direction and magnitude by country, and the results of earlier pilot studies suggest that such differences could depend upon item type as well as by gender and socioeconomic status (Jerrim, 2016).

In anticipation of mode effects, NAEP conducted numerous pilot studies to inform plans for keeping achievement scores linked across paper and digital modes (Bennett et al., 2008; Sandene et al., 2005; Thissen & Norton, 2013; Zhang et al., 2016). To minimize the potential effects of the digital platform on student performance, NAEP (as well as PISA) made changes to some trend items, despite these changes introducing differences in item presentation between modes (Sandene et al., 2005). For example, certain features of items that have been shown to cause mode effects were eliminated or minimized because they inhibited students' abilities to engage with the item, including the need to scroll on the digital interface to see the entire item (Bennett et al., 2008; Bridgeman, Lennon, & Jackenthal, 2003; Pommerich, 2004; Way et al., 2016).

In the United States, two state-led assessment consortia, PARCC (Partnership for the Assessment of Readiness for College and Careers) and SBAC (Smarter Balanced Assessment Consortium) made efforts to minimize threats to device comparability. The consortia conducted small studies to gather information about students' test-taking strategies when using PC and tablet variations (AIR, 2013), as well as feasibility studies to assess the suitability of various devices for test delivery (PARCC, 2017; SBAC, 2014). For example, based on difficulties encountered by students when using on-screen tablet keyboards, the consortia strongly recommended the use of external keyboards when using tablets and also emphasized the importance of providing students practice in using assessment delivery devices prior to testing. To prevent issues related to device and

assessment software reliability, the consortia developed technology guidelines that provide minimum specifications for assessment delivery devices, and created detailed, step-by-step manuals and training modules to help schools prepare devices and administer the computer- and tablet-based assessments.

Based on the existing research and experiences of other large scale assessment programs described above, TIMSS developed its eAssessment system with special consideration given to three potential sources of construct irrelevant variance in student performance on digital assessments:

- The extent of students' familiarity with using computers and tablets for assessment
- The nature of the "response actions" required of students to respond to digital items
- Technical issues associated with computer and tablet devices or assessment software

While the experiences of NAEP, PISA, PARCC, and SBAC helped to inform eTIMSS development, it is critical to consider each unique assessment context when making decisions about salient aspects of digital assessments (APA, 1986; DePascale et al., 2016). TIMSS carefully developed the procedures to transition to eTIMSS in consideration of the unique TIMSS context and the aforementioned challenges.

The Path to eTIMSS

Four data collection efforts during the transition to eTIMSS provide the basis for maintaining trends for eTIMSS countries in TIMSS 2019—the prePilot, Pilot / Item Equivalence Study, Field Test, and the collection of bridge data for the main study.

- The **eTIMSS prePilot** was conducted in September 2016 in three English-speaking countries to try out several newly developed Problem Solving and Inquiry Tasks (PSIs), some items from TIMSS 2015 converted to the eTIMSS format, and the eTIMSS software on PCs and tablets. Conducting the prePilot in English-speaking countries helped avoid issues associated with translation. The results highlighted areas where students encountered difficulties in engaging with the assessment items and informed the converting of paper trend items to the eTIMSS interface.
- The **eTIMSS Pilot / Item Equivalence Study** was administered in May 2017 and data were analyzed through December 2017. The study involved examining the effect of mode of administration (paperTIMSS vs. eTIMSS) on the measurement properties of the trend items that determine the link from TIMSS 2015 to TIMSS 2019 for measuring trends. The same students took the trend items in both paper and digital formats. The results provided a foundation for planning how best to link paperTIMSS and eTIMSS scores for TIMSS 2019 and informed new digital item development. The Pilot / Item Equivalence Study also provided the first large scale opportunity to try out software and application components of the TIMSS eAssessment system.

- A full-scale **Field Test** in March and April 2018 served as a "dress rehearsal" for main data collection with a dual paper and digital delivery system. That is, the countries planning to transition to eTIMSS conducted the Field Test using the full capabilities of the eAssessment system. The paperTIMSS countries also used the eAssessment system to conduct translation and layout verification and print their paper instruments.
- **Bridge data** will be collected concurrently with main data collection in 2019. For countries transitioning to eTIMSS, an additional subsample of students called a "bridge sample" will receive paperTIMSS booklets of the trend items, so that the same sets of trend items, called trend "item blocks," are included in both modes of delivery—digital and paper. The bridge data will be used to form a psychometric link between paperTIMSS and eTIMSS scores.

Description of Dissertation

To investigate the comparability of TIMSS achievement scores between paper and digital modes of administration and describe the foundation for maintaining trend measurements to TIMSS 2019, this dissertation first documents the steps taken to help preserve TIMSS trends from the beginning of eTIMSS development and through the analysis of the data from the eTIMSS Pilot / Item Equivalence Study. The dissertation also extends the analysis of the eTIMSS Pilot / Item Equivalence Study to examine score-level mode effects by students' gender, socioeconomic status, and in relation to students' self-efficacy for using PCs and tablets, or "digital self-efficacy."

With the same students and the same sets of trend items included in both paper and digital modes, the research design and procedures for conducting the eTIMSS Pilot / Item Equivalence Study resulted in data that allowed for an empirical investigation of the equivalence of the TIMSS trend items and scale scores between modes. All students sampled for the eTIMSS Pilot / Item Equivalence Study received one paper-and-pencil test booklet (paperTIMSS) and one booklet equivalent "item block combination" in digital format (eTIMSS)—each containing two blocks of mathematics items and two blocks of science items. The study employed a counterbalanced design, where half the students took eTIMSS first then took paperTIMSS, and the other half took paperTIMSS first then eTIMSS. The design ensured that students were given different items for each test session. In total, 25 countries participated in the eTIMSS Pilot / Item Equivalence Study, with 24 countries at the fourth grade and 13 countries at the eighth grade. National Research Coordinators (NRCs) responsible for overseeing TIMSS in each participating country selected purposive samples of 800 students at each grade that included students with a range of abilities and backgrounds. Sample sizes for analysis included 16,894 fourth grade students and 9,164 eighth grade students.

TIMSS used the results of the eTIMSS Pilot / Item Equivalence Study to inform the methodology to link paperTIMSS and eTIMSS scores and maintain trends in TIMSS 2019. The results determined whether the trend items have equal measurement properties for paperTIMSS and eTIMSS, and thus may be treated as identical in determining the link from TIMSS 2015 to TIMSS 2019. With equal measurement properties across modes, both paperTIMSS and eTIMSS scores can be equated with TIMSS 2015 scores through common item linking. If the trend items do not have equal measurement

properties, a different linking methodology is needed that allows the trend items to have unique item parameters for paperTIMSS and eTIMSS. To ensure the preservation of trends in the latter case, nationally representative bridge samples of students in eTIMSS countries will take the paper trend items, while the usual samples take the items with eTIMSS. Then, common population equating methods can be applied to adjust for the differences in the psychometric properties of the trend items between paperTIMSS and eTIMSS.

Consistent with the current TIMSS design and TIMSS experience, students sampled for the eTIMSS Pilot / Item Equivalence Study at each grade were assessed in both mathematics and science. Thus, each student in the eTIMSS Item Equivalence Database has two achievement estimates for each subject—one for paperTIMSS and one for eTIMSS. The database also includes data collected from a brief student questionnaire for variables measuring student characteristics and their experiences with and attitudes for using computers and tablets.

Further analysis of the eTIMSS Item Equivalence Database investigated the nature of the effect of the new eTIMSS administration on TIMSS achievement scores by student subgroups. The results also contribute to a large body of existing research about predictors of score-level mode effects. As the literature suggests, mixed findings across studies may be related to the different ways computer familiarity and self-efficacy are measured, rapid changes in exposure to technology, and increasing improvements to digital assessment systems and technology (Kingston, 2008; McDonald, 2002; Way et al., 2016).

After constructing and validating an IRT (Rasch) scale of students' digital self-efficacy, the dissertation includes in-depth analysis to examine the relationships of paperTIMSS and eTIMSS scores with background variables of interest, including students' gender, socioeconomic status, and the newly developed scale of digital self-efficacy. Examining these relationships in the TIMSS international context may help explain variation in performance on paperTIMSS and eTIMSS and provide new insight into the impact that digital self-efficacy has on digital test performance beyond that of self-efficacy measures on paper-and-pencil test performance.

Taken together, the results of the eTIMSS Pilot / Item Equivalence Study and the analysis conducted for this dissertation address the following research questions:

1. Do the TIMSS 2019 trend items have equal measurement properties in paper and digital formats?
2. How do item-level mode effects differ by grade, subject, and item type?
3. Without adjusting for mode effects on the trend item parameters, what is the effect of the eTIMSS mode of administration on TIMSS mathematics and science scores?
4. Does the mode of administration differentially affect subgroups of students based on gender, socioeconomic status, and digital self-efficacy?

Following this first introductory chapter to the dissertation, Chapter 2 includes a detailed documentation of the major milestones in developing eTIMSS, beginning with planning for development in October 2014 and through the presentation of the results of the eTIMSS Pilot / Item Equivalence Study and early preparations for the TIMSS 2019 Field Test in December 2017. The chronological account of eTIMSS development

highlights the challenges in developing and conducting a computer-based assessment system, as well as the efforts made by the TIMSS & PIRLS International Study Center to preserve trend measurements.

Chapter 3 provides a description of the research design for collecting data for the eTIMSS Pilot / Item Equivalence Study, including the sample and booklet design and the counterbalanced design for administering both paperTIMSS and eTIMSS to students. Then, the analysis procedures and results of the eTIMSS Pilot / Item Equivalence Study are described for each of three phases of analysis: 1) an a priori analysis of item equivalence based on item characteristics; 2) an item analysis of classical item statistics; and 3) a scale score analysis to examine the effect that the new mode of administration has on TIMSS achievement score estimates. A discussion of the results of the eTIMSS Pilot / Item Equivalence Study describes the psychometric approach TIMSS will use to link paperTIMSS and eTIMSS.

Chapter 4 includes a comprehensive summary of the literature about predictors of mode effects relevant for the TIMSS context and describes the methodology, procedures, and results of the analysis of the eTIMSS Item Equivalence Database. Details about the construction of the IRT scale measuring students' digital self-efficacy are provided in Appendix A.

The fifth and final chapter addresses the research questions explored by the dissertation in light of the results and describes the plans for maintaining TIMSS trend measurements. Suggestions are provided for further research and for further enhancements to the eTIMSS assessment.

Chapter 2: eTIMSS Development History - October 2014–December 2017

This chapter includes a chronological account of the early steps in developing the eTIMSS assessment, beginning in October 2014 and through the presentation of the results of the eTIMSS Pilot / Item Equivalence Study in December 2017. The history focuses on procedures intended to reduce mode effects, as well as obstacles encountered and the rationale for decisions made. The author gathered the information comprising this chapter through a case study approach, relying on internal TIMSS documentation, first-hand experience, and feedback from other individuals involved in development—including staff from the TIMSS & PIRLS International Study Center and IEA Hamburg.

The Executive Directors documented development milestones through regular progress reports presented to TIMSS country representatives at National Research Coordinator (NRC) meetings as well as to the IEA Standing Committee and the IEA General Assembly, the IEA's decision-making authority. The NRCs responsible for overseeing TIMSS in each participating country provided vital feedback about the components of the eAssessment system in conducting the eTIMSS prePilot and the Pilot / Item Equivalence Study, and in preparation for the TIMSS 2019 Field Test.

Overview: Major Development Milestones

In October 2014, Executive Directors of the TIMSS & PIRLS International Study Center, Dr. Ina V.S. Mullis and Dr. Michael O. Martin, unveiled plans for the eTIMSS initiative to the IEA General Assembly at their 55th meeting in Vienna. This group

includes representatives from around the world, appointed by IEA's member countries and institutions, who review IEA's plans and operations. Three years later, the Executive Directors presented the draft results of the eTIMSS Pilot / Item Equivalence Study at the 58th IEA General Assembly meeting in Budapest.

Exhibit 2.1 shows key progress over a three year period from the introduction of eTIMSS to the IEA member countries in October 2014 through reporting the results of the eTIMSS Pilot / Item Equivalence Study in December 2017 to the IEA Technical Executive Group (TEG), who is consulted on all technical aspects of TIMSS.

Exhibit 2.1: eTIMSS Development Milestones, October 2014–December 2017

October 2014	Executive Directors unveiled plans for developing eTIMSS at the 55th Meeting of the IEA General Assembly, Vienna
January 2015	eTIMSS development began for delivery with tablet and stylus: • TIMSS & PIRLS International Study Center began planning to convert existing trend items from paper format to tablet-and-stylus format • Development of Problem Solving and Inquiry Tasks (PSIs) began • IEA Hamburg began building the eAssessment system for eTIMSS
August 2015	TIMSS & PIRLS International Study Center arranged for the American Institutes for Research (AIR) to conduct cognitive labs to inform the conversion of the trend items to tablet-and-stylus format. The labs provided information about: • Feasibility of scrolling • Feasibility of writing with a stylus Results indicated that students had no trouble scrolling, but reported having difficulty using the stylus and said they wrote less than they would have on paper.
October 2015	Executive Directors presented the progress made in developing eTIMSS at the 56th Meeting of the IEA General Assembly, Mexico City: • Approximately 80 percent of trend items were judged as "essentially identical" between paper and tablet formats • The eTIMSS Player was expanded to include keyboard functionality
December 2015	eTIMSS was introduced to the TIMSS National Research Coordinators (NRCs) for delivery with tablet-and-stylus at the TIMSS 2015 7th NRC Meeting, Lisbon: • A timeline was presented for the transition to eTIMSS with requirements to maintain trends, including a doubled sample size to administer both paperTIMSS and eTIMSS in their entirety • IEA presented increased participation fees and other projected costs for administration

Exhibit 2.1: eTIMSS Development Milestones, October 2014–December 2017 **(Continued)**

June 2016	New developments for the transition to eTIMSS were announced at the TIMSS 2015 8th NRC Meeting, Quebec City: • Reliance on tablet-and-stylus technology was abandoned—the eTIMSS platform was extended to include PC and Windows devices as well as a greater variety of Android tablets • A revised approach to maintain trends includes a four step path to eTIMSS—prePilot, Pilot / Item Equivalence Study, Field Test, and Bridge in 2019
October 2016	The eTIMSS prePilot was administered in Australia on tablets, in Canada on PCs, and in Singapore on both tablets and PCs. The results provided information about: • Difficulty setting up devices ahead of testing • Software issues and loss of data during testing • Need for students to use scratch paper for mathematics items due to limitations of the draw tool and stylus • Differences in item presentation between PC and tablets
February 2017	Improvements for eTIMSS based on the results of the eTIMSS prePilot were announced at the joint TIMSS 2015 9th NRC Meeting and the TIMSS 2019 1st NRC Meeting, Hamburg: • Improved instructions/manual for preparing for test administration • Enhanced efforts to improve eTIMSS system reliability • Improvements to eTIMSS user interface for PC
April 2017	Executive Directors provided an overview of eTIMSS development in preparation for the Field Test at the TIMSS 2019 2nd NRC Meeting, Hamburg: • Item development and updating Problem Solving and Inquiry Tasks according to the *TIMSS 2019 Assessment Frameworks* • New technologically enhanced item types include drag & drop, selectable, sortable, and dropdown menu The TIMSS & PIRLS International Study Center's Director of Sampling, Psychometrics, and Data Analysis presented the analysis plans for the Pilot / Item Equivalence Study.
May 2017	The eTIMSS Pilot / Item Equivalence Study was conducted in 24 countries at the fourth grade and in 13 countries at the eighth grade
October 2017	Executive Directors presented the draft results of the eTIMSS Pilot / Item Equivalence Study at the 58th Meeting of the IEA General Assembly, Budapest
December 2017	IEA Technical Executive Group (TEG) reviewed the plans for bridge samples for equating paperTIMSS and eTIMSS at their meeting in Paris

During these three years, TIMSS made substantial progress in meeting the goals for eTIMSS development. However, unforeseen challenges and new insights also led to

several revisions of the initial plans for development to meet measurement goals for TIMSS 2019. The development efforts summarized in Exhibit 2.1 and described in this chapter, including the smaller-scale data collection initiatives informing these efforts, place particular emphasis on the three issues highlighted by the experiences of other assessment programs:

- The extent of students' familiarity with using computers and tablets for assessment
- The nature of the "response actions" required of students to respond to digital items
- Technical issues associated with computer and tablet devices or assessment software

The chapter begins in January 2015 with developing plans for the eAssessment system and converting the trend items from paper to a digital format for delivery on tablets, and concludes with preparing for the TIMSS 2019 Field Test based on the insights gained from the eTIMSS Pilot / Item Equivalence Study.

A Note about Problem Solving and Inquiry Tasks (PSIs)

This dissertation focuses on the challenges of maintaining trends. However, it should be noted that in addition to the milestones listed above, great time and energy was dedicated to a new, innovative initiative for eTIMSS—developing Problem Solving and Inquiry Tasks (PSIs). PSIs are digitally enhanced and interactive, providing more informative assessment in areas of the TIMSS assessment frameworks that were difficult to measure in a traditional paper-and-pencil format. The tasks simulate real world and laboratory situations in mathematics and science and involve students integrating and

applying process skills and content knowledge. The dynamic and animated tasks are visually attractive, which can motivate and engage students. The tasks will collect process data, which can be utilized to analyze students' problem solving and inquiry strategies used to engage with the tasks.

Planning the eAssessment System for TIMSS

To support the development and implementation of eTIMSS, IEA Hamburg began collaborating with the TIMSS & PIRLS International Study Center in January 2015 to plan and develop the eAssessment system for TIMSS. The eAssessment system has five interconnected software and application components:

- **eTIMSS Assessment Builder**—an item-authoring tool for entering digitally-formatted items into the assessment system and assembling the assessment instruments
- **eTIMSS Online Translation System**—for TIMSS country representatives to translate items into the language(s) of instruction, have translations verified by IEA Amsterdam, and have instrument layout verified by the TIMSS & PIRLS International Study Center
- **eTIMSS Player**—for delivering eTIMSS to students, capturing responses, and uploading response data to the IEA server for scoring and processing
- **eTIMSS Online Data Monitor**—for countries to monitor eTIMSS data upload to the IEA server

- **IEA Online Scoring System**—systematically distributes student item responses to constructed response items to trained scorers to score according to scoring guides

Before expanding the eTIMSS platform to accommodate PC devices, original plans for development were limited to tablets equipped with stylus technology, thought to best resemble the student experience on paper. To ensure that the trend items (as well as PSIs) were properly formatted to the digital interface and were capable of being translated into several languages, the Item Builder (part of the eTIMSS Assessment Builder), the Translation System, and the eTIMSS Player were developed concurrently with converting the trend items.

Converting Paper Trend Items to a Tablet-and-Stylus Format

In January 2015, the TIMSS & PIRLS International Study Center began reviewing the 400 trend items from TIMSS 2015 to develop a strategy to convert the paper-formatted items to a tablet-and-stylus format. Research investigating the sources of mode effects has shown mixed results about which item features may enhance or impede students' abilities to provide evidence of the construct (Bridgeman et al., 2003; Johnson & Green, 2006; Pommerich, 2004; Way et al., 2016). Two goals in particular were fundamental in converting the trend items to tablet-and stylus-format:

- Maintain the same presentation of the items across paper and digital modes, as much as possible (Pommerich, 2004)
- Minimize the need for scrolling (Bridgeman et al., 2003; Pommerich, 2004; Way et al., 2016)

In view of maintaining the same presentation of the items between paper and digital modes, TIMSS thought the tablet-and-stylus format best resembled the student test-taking experience with paper and pencil. Initially, TIMSS converted the trend items to the tablet-and-stylus format to allow students to provide constructed answers in ways similar to that of the paper-and-pencil assessment, including using a stylus to show calculations, provide extended written answers, and draw graphs and diagrams.

Tablet-and-Stylus Trend Item Classifications

Based on the initial tablet-based conversions, staff at the TIMSS & PIRLS International Study Center classified the trend items into three categories depending on how similar the presentation was between paper and tablet formats. The results provided insight into the work required to develop the Item Builder and fully adapt the trend items to the eTIMSS interface. Across the fourth and eighth grade assessments, 80 percent of trend items were determined to be "essentially identical"—appearing the same on the tablet as on paper. Approximately 20 percent of items were classified as "readily adaptable"—requiring slight modifications to fit a smaller space, including rearranging or reducing the size of graphics, or requiring some scrolling. Five items at the eighth grade that each comprise two pages on paper were classified as "too big for tablet"—requiring students to scroll to see the entire item.

Designing the User Interface for eTIMSS

In January 2015, the TIMSS & PIRLS International Study Center also began to design a user interface for eTIMSS that facilitated a user-friendly experience and minimized the potential for performance differences between paper and digital

assessment delivery modes. The user interface includes the elements of eTIMSS that students see and interact with on the tablet or PC:

- Physical layout of the screen (portrait vs. landscape)
- The means for moving within and across item screens (navigation)
- Assessment tools (e.g., ruler, calculator) and "response actions" required of students to respond to items (e.g., clicking, using a number pad to enter numerical answers)

Because some students from the diverse TIMSS countries may have had little prior experience with using digital devices (Skryabin, Zhang, Liu, & Zhang, 2015), staff at the TIMSS & PIRLS International Study Center designed the eTIMSS user interface to be universally easy and intuitive to navigate. The overall goal in designing a user interface is for students to attend to test content and not be preoccupied with the mechanics of using the Player (Parshall, Spray, Kalohn, & Davey, 2002; Pommerich, 2004). It should be made clear to students what part of the screen to attend to, how to navigate within and across screens, and how to respond to items.

The TIMSS & PIRLS International Study Center developed the user interface design in consideration of research-based principles for designing multimedia instruction and educational games that are associated with improved cognitive performance and greater student enjoyment (Falloon, 2013; Mayer, 2009; 2014). For example, any buttons used to navigate within eTIMSS, activate tools, or respond to items should function the same way throughout the assessment. There is no extraneous or distracting material on the interface, with essential material highlighted on the screen. Assessment tools appear highlighted along the bottom navigation bar when they are

active for a particular item, with inactive tools appearing "greyed out." Item screens that students have not viewed appear bright-colored in the left-hand progress bar for students to revisit and respond within the given time, with buttons for visited item screens darkened.

Developing the eTIMSS user interface involved an iterative development process, focusing on isolated aspects during each phase (for examples, see Pommerich, 2004; Way et al., 2016). During the early stages of eTIMSS development, staff at the TIMSS & PIRLS International Study Center designed the user interface in conjunction with the conversion of paper trend items to digital format.

Layout

A portrait (vertical) tablet layout better resembles the paper-and-pencil experience, and TIMSS decided to have a portrait layout instead of a landscape (horizontal) layout based on a careful analysis of the height of the trend items in their paper formats. TIMSS developed the trend items for paper booklets in measurement lengths called "bar slots," based on the length of the vertical formatting bar applied to the left side of each item. One bar slot is approximately 3.5 centimeters tall and there are six bar slots available on each paper booklet page. A careful analysis of all of the trend items indicated that items of four bar slots or less could be converted to a digital portrait format without modifying the layout and only 20 percent of items at each grade were five or six bar slots. Therefore, to preserve the layout of the trend items, the eTIMSS interface uses a portrait layout.

Navigation

The eTIMSS interface includes three easy-to-use navigation panels. The TIMSS & PIRLS International Study Center designed these panels to take up as little screen space as possible to accommodate a larger space for item content. On the left side of the screen, a vertical navigation bar, or "progress bar," allows students to track their progress toward completing the assessment and to skip to any item screen with sequentially numbered buttons. After a student has visited an item screen, the buttons change color. If item content does not fit on a single screen, students use the scroll bar is on the right side of the interface to navigate vertically within an item screen. On the bottom of the screen, a navigation panel displays buttons to move forward to the next screen or backward to the previous screen. This bottom panel also includes buttons to activate assessment tools used to respond to items (e.g., ruler, calculator).

Assessment Tools and Response Actions

The way in which students respond to digital versions of the trend items—called "response actions" (Parshall, Davey, & Pashley, 2000)—is dependent on the item layout and the type of student responses specified in the scoring guides. Among the TIMSS trend items, multiple-choice items require students to select answer option bubbles, while constructed response items have varied response requirements. For example, trend items requiring numerical answers were programmed to activate a number pad equipped with the digits 0 through 9, a negative sign (-), a decimal point (.), and a fraction format when the input space became active. Based on the trend items, the assessment tools originally included for the AIR Cognitive Labs included a draw tool and eraser.

AIR Cognitive Labs - August–September 2015

By mid-2015, converting the trend items, developing the Item Builder, and designing the user interface for eTIMSS continued apace to the point that several blocks of items were completed. Under the direction of the TIMSS & PIRLS International Study Center, the American Institutes for Research (AIR) used these blocks to administer cognitive interviews, called cognitive "labs," in August and September of 2015. The TIMSS & PIRLS International Study Center provided AIR with Samsung Galaxy Tab A tablets with styluses that were equipped with a prototype of the eTIMSS Player containing a subset of trend items and several PSIs. Based on research questions related to the functionality of the stylus and the feasibility of scrolling, interview protocols were developed that incorporate a think aloud aspect and retrospective aspect for a small subsample of students (AIR, 2015).

During the interviews, students explained their thoughts while engaging with the items and provided insights into how both the item format and the eTIMSS user interface could be improved. Overall, students were comfortable with scrolling. However, about half the students at each grade struggled with using a stylus to show calculations, provide written explanations, and draw diagrams, and reported that they wrote less than they would have on paper.

Based on these results, the eTIMSS Player, eTIMSS user interface, and digital trend items were updated to allow students to use the on-screen keyboard to answer constructed response items, rather than be limited to a stylus. These changes were presented along with the tablet-and-stylus trend item classifications at the 56th meeting of the IEA General Assembly in October 2015 in Mexico City, Mexico.

Feasibility Constraints: Accommodating Country Diversity

When developing eTIMSS, TIMSS made considerations for the diverse national contexts of TIMSS country participants as well as for the capabilities of IEA Hamburg to build a computer-based assessment system appropriate for TIMSS. Original plans for countries to transition to eTIMSS and maintain trend measurements included a doubled sample size requirement to administer both paperTIMSS and eTIMSS in their entirety for TIMSS 2019. In addition, eTIMSS only allowed tablets equipped with a stylus. These plans were announced in December 2015 at the TIMSS 2015 7th National Research Coordinator Meeting in Lisbon, Portugal, along with projected costs for administration. However, concerns arose over the increased burden for test administration, including increased participation fees, as well as initial investment costs to purchase devices.

To encourage more countries to participate in eTIMSS, the eTIMSS platform was extended to include PC and Windows and a greater variety of Android tablet devices. Additionally, TIMSS revised the size of the paper bridge sample to reduce the burden for countries transitioning to eTIMSS. Rather than requiring a second, full sample of students to complete the full set of paper booklets during the TIMSS 2019 data collection, only a smaller additional sample of 1,500 students will be required to complete paper booklets, and the paper booklets only will contain the blocks of trend items. These modifications for eTIMSS were announced in June 2016 at the TIMSS 2015 8th National Research Coordinator Meeting in Quebec City, Canada.

These decisions had implications for other aspects of development. IEA Hamburg had to make the eTIMSS Player delivery software compatible with additional device platforms. In addition, the TIMSS & PIRLS International Study Center re-designed the

eTIMSS user interface and some trend items to minimize the potential influence of students' prior experience with PCs and tablets on digital test performance, and to minimize issues of device comparability (Davis et al., 2017; Way et al., 2016).

Although eTIMSS may be more efficient than paperTIMSS for conducting TIMSS in the long-term, the development and implementation of eTIMSS called for additional up-front costs for countries to prepare for new testing procedures. This included investing in devices for test delivery and ensuring schools have adequate IT infrastructure to conduct the assessment.

Because in some countries, schools are not equipped with PCs or tablets, countries may need to purchase or rent digital devices for test delivery. In some countries, distributing devices from a central location could require expensive shipments or travel to remote areas. Renting digital devices would require that the schools keep the devices in fair condition for return. Schools also need physical space for test administration, including appropriate buildings and facilities, possibly requiring access to electrical outlets to charge devices. Portable devices may also need re-charging between testing sessions. In addition, although the eTIMSS Player does not require an Internet connection for test delivery—a connection is required to upload the data to the IEA server. This may introduce additional complexities for countries where Internet access is limited.

To minimize initial investment costs for eTIMSS countries who have already purchased devices for other computer-based assessments, eTIMSS can be delivered on both PCs and tablets—as long as the tablets run on the latest Android operating system. With a variety of devices now deemed acceptable, TIMSS put in place some limitations regarding device specifications to ensure that eTIMSS software will run properly. For

example, older tablets may not have the capacity for the latest Android operating system, and may cause the assessment to run slow or response data to be lost. Therefore, countries may need to update device software accordingly.

eTIMSS prePilot - September–October 2016

Research has shown that desktop and laptop PC devices have similar "form factors," which describe the way students use the device when engaging with the test (Davis et al., 2017). Student test performance using desktop and laptop PCs has shown to be relatively comparable, with differences in performance attributed to technical issues occurring with some devices and not others (Davis et al., 2017; Sandene et al., 2005). However, there are greater differences in the response actions required of students on PCs compared to touch-screen tablets. First, tablet screens can be smaller than PC screens, which could limit the amount of information students can see on the screen at one time. Also, inputting responses may be more or less difficult with a mouse, finger, or stylus, depending on the response action required (Strain-Seymour et al., 2013).

In light of this research, the TIMSS & PIRLS International Study Center made efforts to improve the eTIMSS interface to make the eTIMSS experience as similar as possible across variations in PC and tablet devices. For example, any buttons used to navigate within eTIMSS, activate tools, or respond to items require approximately the same action by students on both PCs and tablets. Additionally, the TIMSS & PIRLS International Study Center added a directions module to the eTIMSS assessment. Through practice, the directions module provided at the beginning of the assessment allows students to become familiar with navigating the eTIMSS Player screens and the

assessment tools, and illustrates the response actions required for responding to the different types of items. The results of the eTIMSS prePilot further informed development efforts to reduce device-related performance differences.

The eTIMSS prePilot was conducted in September and October of 2016 to try the eTIMSS Player on both PCs and tablets in a classroom setting, including the newly added capability to use a keyboard to answer constructed response items. The digital prePilot instruments included a subset of TIMSS mathematics and science items from the 2015 cycle that were released for restricted use as well as four PSIs. The prePilot was conducted at fourth and eighth grades in three English-speaking countries with experience in conducting digital assessments: Australia, Canada, and Singapore. NRCs from Australia and Singapore each provided the TIMSS & PIRLS International Study Center with a detailed report on the prePilot that included feedback about setting up the devices for testing, the reactions and difficulties of students, and issues that occurred with digital devices during testing or when uploading the data.

In Australia, the Australian Council for Educational Research (ACER) administered the prePilot on Samsung Galaxy Tab A tablets in two classes of fourth grade students and two classes of eighth grade students. Overall, students reported enjoying taking a test on a tablet. Students were engaged with the test content and navigated within the eTIMSS interface with no problems. However, students reported difficulty using the stylus to write and draw. The ACER test administrators also reported some difficulties with the tablet Player, including students being logged out of the program and subsequently losing some response data.

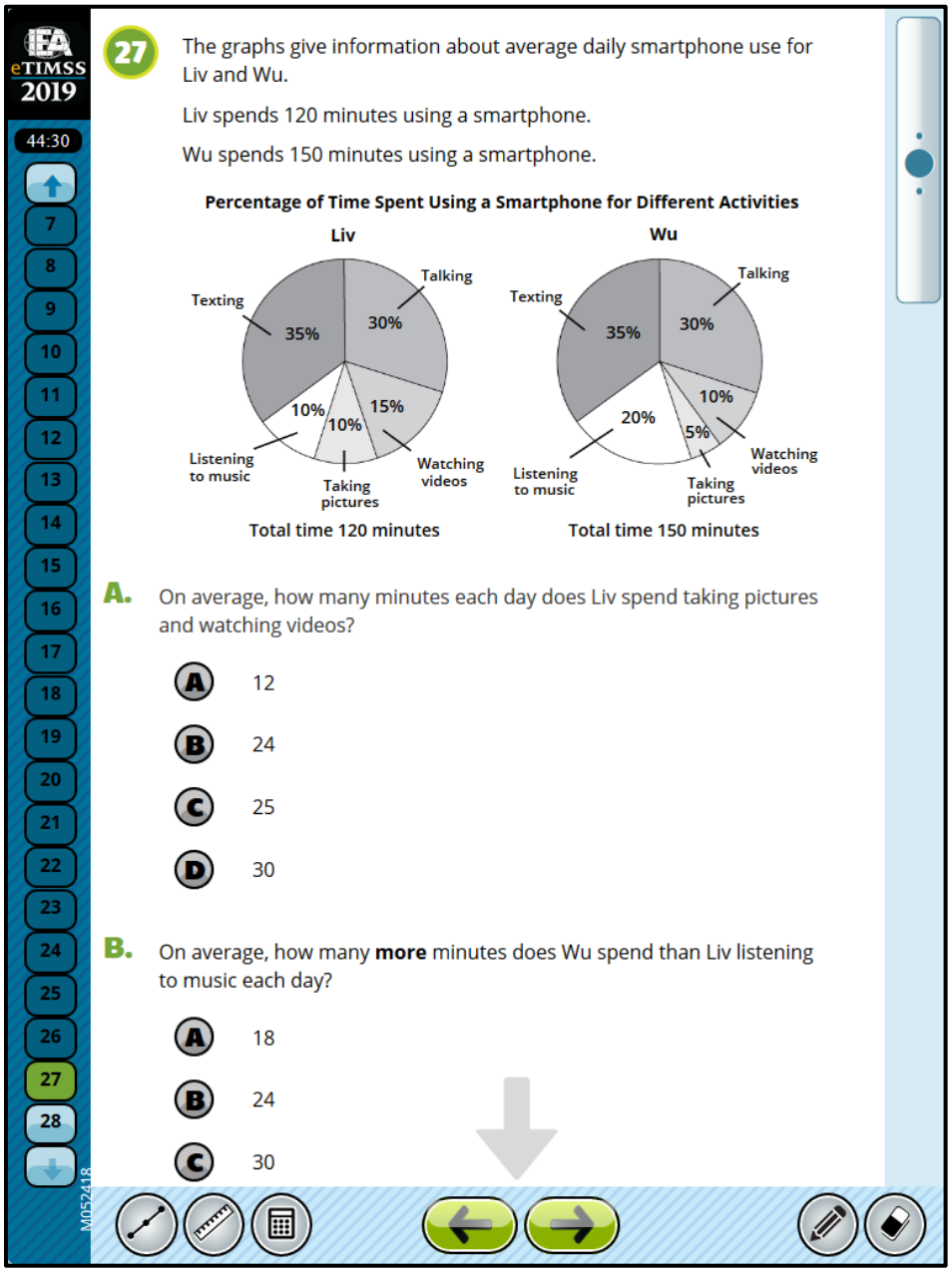
Singapore administered the prePilot in two schools per grade, with two classes per school. Students were randomly assigned to use a PC or tablet, stratified by gender. While results indicated that students were engaged with the items, some students reported experiencing difficulty navigating from screen to screen and with using some assessment tools on a tablet, including the draw tool and eraser tool as well as the number pad. Students took less time to answer science items compared to mathematics items, which reportedly required more effort by students to input answers using a stylus. Students often relied on scratch paper to solve mathematics items, and became frustrated using the number pad, stating they were unfamiliar with the layout of the numbers, which was different than a typical number pad on a PC keyboard.

Test administrators from Singapore also reported instances of the eTIMSS Player software "crashing" during administration. In addition, students were able to exit the eTIMSS Player on tablets during the test session, students using tablets were distracted by other applications available and did not remain engaged with the test. This type of distraction did not occur on PCs, where a lock feature was active.

The student response data collected for the prePilot showed that substantial differences in item appearance between tablets and PCs were possibly disadvantageous to students using PCs. While text and images appeared larger on a PC, the larger screen size had a negative impact on the layout of items, with more scrolling required to see the entire item. To fix this, the TIMSS & PIRLS International Study Center modified the eTIMSS interface following the prePilot so that the items have a maximum width and a similar width to length ratio on any compatible device. For items that required scrolling, an arrow was programmed to appear at the bottom of the screen to indicate to students

that there is more content. Exhibit 2.2 shows an example item screen from the eTIMSS Pilot / Item Equivalence Study delivered on a PC, showing all features of the eTIMSS interface present for the study. The screen illustrates the width to length ratio, similar to that of a tablet, as well as the added arrow to indicate that scrolling is needed.

Exhibit 2.2: Example Item Screen from the eTIMSS Pilot / Item Equivalence Study



Source: eTIMSS 2019 Pilot / Item Equivalence Study, Eighth Grade
Note: Trend item is confidential. Do not cite or circulate.

In response to countries' feedback from conducting the eTIMSS prePilot, as well as the need to accommodate device variations, the TIMSS & PIRLS International Study Center also made changes to the number pad, the eTIMSS Player software, and manuals for setting up devices for testing and conducting the test session in advance of the eTIMSS Pilot / Item Equivalence Study. Additionally, TIMSS decided to abandon its reliance on stylus technology for eTIMSS following administration of the Pilot / Item Equivalence Study due to students' frustration and limitations of the IEA's Scoring System at the time of the study.

The Executive Directors announced these changes for eTIMSS to National Research Coordinators (NRCs) at the joint TIMSS 2015 9th NRC meeting and the TIMSS 2019 1st NRC meeting in Hamburg, Germany in February 2017. Around this time, IEA Hamburg began working to improve the reliability of the eTIMSS Player software to prevent crashing and loss of data. Feedback about the issues encountered in setting up devices and solving software issues during the test session were informative for writing more detailed instructions in the manuals for the eTIMSS Pilot / Item Equivalence Study. To ensure the eTIMSS systems can fully operate on whatever device(s) countries choose to use, IEA Hamburg developed the eTIMSS SystemCheck program in advance of the eTIMSS Pilot / Item Equivalence Study—an application for PCs (and on for tablets for the TIMSS 2019 Field Test)—to check whether they fulfill the minimum requirements for running eTIMSS.

Unfortunately, re-designing the software meant that IEA Hamburg had to delay the eTIMSS Pilot / Item Equivalence Study. This resulted in several countries having to

drop out at the eighth grade due to conflict with other high stakes tests occurring at the end of the school year.

eTIMSS Pilot / Item Equivalence Study - January–June 2017

Formal preparations for the eTIMSS Pilot / Item Equivalence Study began in January 2017. Preparations by the TIMSS & PIRLS International Study Center and IEA Hamburg included refining and testing the digital trend items, testing the eTIMSS Player, and completing the Online Translation System for use by country representatives. Additionally, IEA Hamburg prepared its Online Scoring System to score constructed response items for eTIMSS and the IEA Online Data Monitor for country representatives to monitor eTIMSS data upload.

National Research Coordinators (NRCs) from each participating country were responsible for overseeing the implementation of the eTIMSS Pilot / Item Equivalence Study in their countries. After an overview of the eTIMSS Pilot / Item Equivalence Study, the following sections outline the chronological phases of conducting the study. Each section describes procedures for the study, along with any challenges encountered and changes made to eTIMSS made based on those challenges.

Overview of the eTIMSS Pilot / Item Equivalence Study

The eTIMSS Pilot / Item Equivalence Study was conducted in 25 countries—24 at the fourth grade and 13 at the eighth grade—that are planning to transition to eTIMSS for the TIMSS 2019 assessment. The study served two primary purposes:

- Examine the effect of mode of administration (paperTIMSS or eTIMSS) on the measurement properties of the trend items that determine the link from TIMSS 2015 to TIMSS 2019 for measuring trends
- Try out components of the eTIMSS Assessment System—including the eTIMSS Translation System, Player, Scoring System, and Data Monitor

Primarily in May 2017, paper and digital achievement instruments consisting of the same trend items were administered to fourth and eighth grade students. The items are the complete set of trend items from TIMSS 2015—187 at the fourth grade and 232 at the eighth grade. Following the TIMSS 2015 assessment, eight of the mathematics item blocks and eight of the science item blocks were secured to use in 2019 to measure trends at both the fourth grade and at the and eighth grade. The items were in eight different booklets designed to last 72 minutes at the fourth grade and 90 minutes at the eighth grade.

The TIMSS & PIRLS International Study Center designed the eTIMSS Pilot / Item Equivalence Study to examine the equivalence of the TIMSS trend items between paper and digital administration modes and to inform procedures for linking paperTIMSS and eTIMSS scores for TIMSS 2019. Each student sampled received one paperTIMSS booklet and one eTIMSS "item block combination" with a different set of items than those in the paper booklets.

The second major purpose of the eTIMSS Pilot / Item Equivalence Study was to provide countries practice in using the TIMSS eAssessment system components and in conducting a computer-based assessment on a relatively large scale. For many participating countries, the eTIMSS Pilot / Item Equivalence Study provided first time

experience in conducting a computer-based assessment. The experiences of countries and the feedback from Test Administrators and National Research Coordinators were crucial in informing further development of the eTIMSS 2019 assessment in preparation for the Field Test.

Each phase of the eTIMSS Pilot / Item Equivalence Study had a detailed manual explaining how to accomplish each step. The TIMSS & PIRLS International Study Center provided countries with the following manuals:

- Preparing Paper Booklets (January 2017)
- Tracking Class/Student Participation (February 2017)
- Adding Trend Translations to the eTIMSS Translation System (March 2017)
- Preparing Computers for eTIMSS (March 2017)
- Test Administrator Manuals for paperTIMSS and eTIMSS (April 2017)
- Scoring the Constructed Response Items (May 2017)
- Entering and Submitting the Tracking and paperTIMSS Data (May 2017)

Preparing Paper Booklets

For assembling the paper booklets just as in TIMSS 2015, the TIMSS & PIRLS International Study Center provided Adobe InDesign booklet production files to countries so they could recreate booklets using their same translated trend blocks as used in TIMSS 2015. Countries also needed to copy in the new directions for the eTIMSS Pilot / Item Equivalence Study and put the blocks in the correct positions within the booklets. The manual on "Preparing Paper Booklets" guided countries through the process, highlighting similarities and differences compared to the process followed for TIMSS 2015. The manual instructed countries to thoroughly check the assembled booklets before printing.

Enough booklets had to be printed and distributed from national centers to account for any misprints or loss of booklets during distribution and have replacements prepared as necessary.

Tracking Class/Student Participation

The TIMSS & PIRLS International Study Center provided Student Tracking Forms and labels to countries to ensure that each student received the correct paper booklets and digital item block combinations—so that students did not receive the same items in both delivery modes. A Microsoft Excel template determined the order of administration for each class—half of the classes did paperTIMSS first, and half did eTIMSS first. The forms assigned a booklet and item block combination to each student with a password to enable tracking student participation. The Student Tracking Forms also tracked the device used by students for eTIMSS delivery and included birthdate and gender information. A Microsoft Word template produced labels for paper booklets from the information in the Student Tracking Forms.

Exhibit 2.3 shows an example Student Tracking Form for a class. To create the Student Tracking Forms, countries translated the template and labels into the language(s) of the assessment. Then, a list of schools and classes to be assessed was specified, as well as the language(s) of the assessment instruments. Completing the Student Tracking Forms for each class required that country representatives choose the school from the specified list using a dropdown menu, select the number of classes to be assessed, and specify the number of students in each class within the school. Then, forms were automatically created for each class in a school, with half of classes receiving paperTIMSS first and half receiving eTIMSS first (indicated in the "Administration

Sequence" field in the form). The forms populated with student identification numbers and assigned paperTIMSS booklets and eTIMSS passwords (which indicated the eTIMSS item block combination) for each student.

Additional form fields were by completed manually by school personnel with the testing dates and times as well as each student's name or ID number (column 1), birthdate (column 3), gender (column 4), and participation status (column 8). Test administrators or School Coordinators specified each student's participation status at the test session as "P" if the student used a PC for eTIMSS, "T" if a student used a tablet, or "A" if the student was absent on the day of testing.

Exhibit 2.3: Example Student Tracking Form for the eTIMSS Pilot / Item Equivalence Study

eTIMSS 2019 Pilot/ IES Student Tracking Form - Grade 4

Administration Sequence:	First paperTIMSS then eTIMSS

paperTIMSS Date DD / MM		paperTIMSS Time HH : MM		eTIMSS Date DD/MM		eTIMSS Time HH:MM	
[a]	[b]		[c]	[d]	[e]	[f]	

School Name	Country Name	School ID	Class ID	Class	Language
EXAMPLE SCHOOL	EXAMPLE COUNTRY	0004	00040		English
❶	❷	❸	❹ ❺	❻	❼	❽

Student Name or Number	Student ID	Birth (Month / Year)		Gender	paperTIMSS Booklet	paperTIMSS Checksum (for data entry)	eTIMSS Password	Participation Status	
		MM	YYYY					paperTIMSS Session	eTIMSS Session
1	00040101				5	44324	27223		
2	00040102				6	44492	28254		
3	00040103				8	44920	22269		

Preparing Computers for eTIMSS

The manual on "Preparing Computers for eTIMSS" instructed countries how to select and prepare devices for the eTIMSS Pilot / Item Equivalence, including technical specifications for devices and steps for testing and troubleshooting issues with the eTIMSS Player in advance of testing. Exhibit 2.4 presents the PC and tablet device specifications on which eTIMSS could be delivered for the Pilot / Item Equivalence Study. Countries were encouraged to choose the devices based on students' familiarity with using them in the classroom.

Exhibit 2.4: eTIMSS 2019 Pilot / Item Equivalence Study Minimum Device Requirements*

Specifications	PC	Tablet**
Operating systems	Windows XP (SP3) Windows Vista Windows 7, 8, 10	Android 5.0.2
Screen	Resolution: 1366 × 768 pixels or 1280 × 800 pixels	9.7" XGA LCD Resolution: 1024 × 768 pixels
Processor	1.5 GHz	1.2 GHz
Memory	1 GB	2 GB
Other	USB Port 2.0 System Memory of 2 GB	16 GB Storage 1 GB of available space

* PCs and tablets should meet or exceed specifications.
** Tablet requirements for the eTIMSS Pilot / Item Equivalence Study were based on the specifications of "standard" Samsung Galaxy tablets—tested and approved by IEA Hamburg and the TIMSS & PIRLS International Study Center.

IEA Hamburg developed the eTIMSS SystemCheck program to run on PCs via USB stick to determine whether they are capable of running the eTIMSS Player. Countries could check compatibility of tablet devices for the eTIMSS Pilot / Item Equivalence Study with a test version of the eTIMSS Player, which was provided to each

country by IEA Hamburg through the IEA FTP server. The program checks the screen resolution, processor speed, and available memory of the device.

The “Preparing Computers for eTIMSS" manual provided step-by-step instructions for using the SystemCheck program and the tablet test eTIMSS Player. IEA Hamburg also held individual country sessions at National Research Coordinators’ meetings, where National Research Coordinators could check devices and learn about the technology requirements for conducting eTIMSS.

The manual also provided instructions on how to deal with common issues encountered when running the eTIMSS Player software. Fortunately, the instructions mostly covered the difficulties experienced by countries in setting up eTIMSS for the eTIMSS Pilot / Item Equivalence Study. For example, some anti-virus software may prevent the eTIMSS Player software from launching, a computer may freeze or crash during eTIMSS administration, or text and graphics may not appear correctly. The manual described steps to follow in the case that each of these issues were encountered. However, neither the TIMSS & PIRLS International Study Center nor IEA Hamburg anticipated some other issues experienced. For example, one country reported that school computers required students to logon to a network before using them, causing unexpected delays during test sessions.

Adding Trend Translations to the eTIMSS Translation System

To translate the digitally formatted trend items into the language of instruction, country representatives copied the exact translations used in the TIMSS 2015 paper item blocks into the eTIMSS Online Translation System, so that the digital item text matched the paper booklets. Countries provided additional translations for the eTIMSS directions

module as well as for the student questionnaires administered for the Pilot / Item Equivalence Study. IEA Hamburg used the translations to prepare the eTIMSS Player for each country to install onto USB sticks (for PCs) or tablets. The "Adding Trend Translations to the eTIMSS Translation System" manual also instructed countries on preparing USB sticks with the national eTIMSS Players and/or installing the eTIMSS Player on tablets.

The eTIMSS Translation System presented the international English translations for each item, with text fields provided for countries to paste the trend translations for item stem(s), answer options, answer lines and units, and labels for diagrams. Countries could choose a style for each multiple-choice item to specify numerals, Greek letters, or Hindi letters for answer options, instead of the A, B, C, D letter answer options. Access was provided to an advanced image editor (SVG editor) to translate and format diagram labels, and items were available for preview at any time during the translation process. Countries could also add comments for components of items to communicate issues to IEA Hamburg.

Frustrations arose over the Translation System's limited ability for unique national adaptations, requiring extra work by staff at IEA Hamburg to assist countries in preparing instruments for the eTIMSS Pilot / Item Equivalence Study. First, countries could not implement adaptations to answer option buttons described above to all items at one time, requiring that countries change this feature individually for each item. Also, the system was not yet capable of adapting the digital items into right-to-left formatted languages. In paper booklets, right-to-left language formatting often involves horizontally

flipping images. However, countries were not able to flip images without assistance from IEA Hamburg, and many adaptation needs differed by country.

At the time of the eTIMSS Pilot / Item Equivalence Study, the Translation System was unable to accommodate unique characters or mathematics symbols used by some countries. Many unique variations of mathematical symbols were unrecognizable by the system, such as a half character space and special symbols for graphs and arithmetic. Also, the number pad used by students to enter numerical answers was not capable of national adaptations. Countries that use a decimal comma instead of a decimal point could not substitute the appropriate symbol for their students. This required the TIMSS & PIRLS International Study Center to write some additional instructions in the Test Administrator Manuals to avoid confusion by students. Countries were told to adapt these manuals where necessary, instructing students how to use the number pad.

Test Administrator Manuals for paperTIMSS and eTIMSS

The TIMSS & PIRLS International Study Center developed new administration procedures for the eTIMSS Pilot / Item Equivalence Study that included two modes of administration and two separate testing sessions. Therefore, countries received two Test Administrator Manuals—one for paperTIMSS and one for eTIMSS. For each mode of administration, there were instructions for countries to prepare instruments, administer the test, and follow the necessary steps for IEA Hamburg to receive student response data on the IEA server.

The first parts of the eTIMSS Pilot / Item Equivalence Study Test Administrator Manual described how to prepare PCs and/or tablets for the test session. Device preparation proved to be a labor-intensive process for some countries. Because of

security concerns about TIMSS' confidential trend materials as well as personal identifying information that may be included in the data, "cloud" features could not be used that allow the software to be run remotely from the IEA server. Countries used USB sticks to install the eTIMSS Player on each PC and run the SystemCheck program. Installing the Player on tablets required that tablets connect to a PC with internet access to copy the application files onto the device for installation.

The eTIMSS Pilot / Item Equivalence Study Test Administrator Manuals included instructions for using the Student Tracking Form to assign and distribute instruments to students, logging students into the eTIMSS Player, and planning for each testing session. There was a unique Test Administration Script for paperTIMSS and eTIMSS, respectively, as well as instructions for guiding students through the directions module for eTIMSS. The directions module placed particular emphasis on describing how the navigation buttons work and how to respond to item types for which the response actions may be unfamiliar to students. Based on the results of the prePilot, it was especially important that Test Administrators were prepared to deal with any issues encountered during the eTIMSS test session to prevent disengaging students or losing data. Test Administrators needed to be ready at all times to help a student use the eTIMSS interface or to troubleshoot technical issues.

For eTIMSS administration, the eTIMSS Player software saved student item responses, including drawings, directly on the PC and tablet devices used by students. However, uploading data to the IEA server required that the devices access the Internet. In some cases, countries had to bring USB sticks and tablets to locations with Internet access to upload the data. Countries were able to monitor the eTIMSS data submission

via the eTIMSS Online Data Monitor to ensure data were captured and uploaded correctly.

Country representatives who participated in the Pilot / Item Equivalence Study reported that distributing the eTIMSS assessment was mostly completed with ease. However, some countries received reports from schools having difficulty setting up devices for eTIMSS, and staff from national centers had to visit schools to help them prepare. Additionally, many countries complained that the tablet version of the eTIMSS Player did not indicate whether data upload was successful, at times resulting in lost data or duplicate student records if data were uploaded twice. Although countries could monitor data upload through the IEA Online Data Monitor, it was not clear which devices the records were coming from.

Scoring the Constructed Response Items

Using the same scoring guides used for TIMSS 2015, constructed response items were scored by the same scorers in both paper and digital formats to ensure consistency of scoring. To control for possible scoring bias that could occur between paper and digital modes (Horkay et al., 2006; Russell, 2002), scorers blended scoring for the two modes, alternating between the two.

Scoring for paperTIMSS was done on paper, as usual, which is often a labor-intensive process. This process involves appropriately distributing booklets to scorers, later collecting the booklets, and manually entering or scanning the data into the IEA Data Management Expert (DME) software. For eTIMSS, IEA Hamburg refined their scoring system for the Pilot / Item Equivalence Study. The Online Scoring System allows for assigning item responses to scorers according to a systematic plan. The system shows

scorers screen captures of students' responses with the scoring guide for scorers to input scores into the system, directly onto the IEA server.

Scoring for eTIMSS is a much more efficient process and over half the constructed response items will be machine scored for the TIMSS 2019 Field Test. These include items with numerical answers, where students use a number pad to input responses, among other item types. This will greatly reduce the hand-scoring burden for countries and increase overall efficiency of data collection in the future. In addition, the Online Scoring System allows scoring supervisors to easily monitor scorers for accuracy and reliability. However, for the Pilot / Item Equivalence Study, all digital items were hand scored through the IEA Scoring System to test the reliability of the system.

Preparing for scoring responses from the eTIMSS Pilot / Item Equivalence Study involved planning by countries to ensure an adequate number of properly-trained scoring staff for scoring the TIMSS trend items. Countries used paper scoring training materials from TIMSS 2015 to train scorers, and TIMSS input example and practice papers into the Scoring System for scorers to practice scoring digital item responses.

Additional unforeseen technical preparations and procedures were required to prepare some trend items for scoring. The eTIMSS Player and IEA Scoring System were not fully prepared for capturing student drawings or writing on "canvas" item types, and so some constructed response items (13 at the fourth grade and 9 at the eighth grade) required special instructions for scorers. Unfortunately, many responses to canvas items could not be scored at all. At the time of the eTIMSS Pilot / Item Equivalence Study, the IEA's Scoring System could not capture written or drawn responses such that the system could re-create responses on top of the item screen and be legible by scorers. Drawings

were misaligned with the screen and often were uninterpretable. Therefore, items were not scored when a drawing had to be precise.

Entering and Submitting the Tracking and paperTIMSS Data

The "Entering and Submitting the Tracking and paperTIMSS Data" manual guided country representatives and data managers through the procedures for entering and submitting paperTIMSS data for the eTIMSS Pilot / Item Equivalence Study. The eTIMSS Player software uploaded data for the computer and tablet delivery directly to the IEA eTIMSS server. For paper items, countries manually entered or scanned data into the IEA Data Management Expert (DME) software. The DME software allows countries to enter, manage, and organize data, detect and repair data errors, and conduct scoring reliability for constructed response items. After passing quality control checks, countries submit the data to IEA Hamburg for further processing and quality control.

Similar to procedures followed for data entry in TIMSS 2015 (Johansone, 2016), the manual provided by the TIMSS & PIRLS International Study Center described how to:

- Install and prepare the DME software
- Enter the data and code for missing responses
- Import tracking information from the Student Tracking Forms
- Conduct quality control, including checking for duplicate student identification codes
- Export the data to IEA Hamburg

Along with the DME software, countries received codebooks that described the properties and layout of the variables form the paper booklets as well as the variables

from the Student Tracking Forms. There was one codebook template for fourth grade and another for eighth grade. Countries used these templates to create the datasets in the DME software to input the paper data.

The coding of missing data was very important to ensure that student responses are scored properly by macros written by the TIMSS & PIRLS International Study Center and result in accurate item statistics. Missing responses were coded as "not administered" when items were misprinted, damaged, or missing, or if students for some reason did not have a chance to read the item. Responses were coded "omitted," when the item was administered but not answered. Some omitted responses were later computer-scored as "not reached," indicating that a student did not have enough time to finish a section of the test booklet.

IEA Hamburg developed the Student Tracking Forms in a template that allowed the student tracking data to be imported into a DME database. These data were especially important for the eTIMSS Pilot / Item Equivalence Study to ensure student identification codes matched up across paperTIMSS and eTIMSS testing sessions. For each student, the data from the Student Tracking Forms indicated the assigned paperTIMSS booklet and the eTIMSS item block combination, the order of administration for the student's class, the device used for eTIMSS delivery, and information about the student's birthdate and gender. These data are included in the final eTIMSS Pilot / Item Equivalence Database, described in Chapter 4.

Feedback from Country Participants

Throughout the process of conducting the eTIMSS Pilot / Item Equivalence Study, country representatives provided feedback to IEA Hamburg and the TIMSS &

PIRLS International Study Center via e-mail about the eTIMSS Online Translation System, eTIMSS Player, and IEA Online Scoring System. After countries submitted their data for processing, IEA Hamburg collected explicit feedback about their experiences using the eTIMSS Online Translation System, distributing and setting up the eTIMSS assessment in schools, and the performance of the eTIMSS Player. In addition, almost all participating countries provided documentation of all differences between paper and digital versions of trend items. IEA Hamburg used this country documentation to improve the eAssessment system components and the TIMSS & PIRLS International Study Center referenced the documentation during item review for the eTIMSS Pilot / Item Equivalence Study to investigate any country-specific or general item issues (see Chapter 3).

Preparing for the Field Test - June 2017–April 2018

A major purpose of the eTIMSS Pilot / Item Equivalence Study was to try out the eTIMSS systems on a large scale and provide countries with practice in conducting computer-based assessments. National Research Coordinators and Test Administrators provided extremely informative feedback during and after the eTIMSS Pilot / Item Equivalence Study. Based on this feedback and the results of the eTIMSS Pilot / Item Equivalence Study presented in Chapter 3, the TIMSS & PIRLS International Study Center and IEA Hamburg made refinements to the eTIMSS user interface, newly developed items, and all eAssessment system components in preparation for the TIMSS 2019 Field Test. Country feedback was especially critical for writing manuals for conducting the Field Test. These preparations formally began in June 2017.

Updating the eTIMSS Online Translation System

Following the eTIMSS Pilot / Item Equivalence Study, preparing for the Field Test involved improving the software components of the eAssessment system so that countries are able to implement standardized testing procedures appropriate for their cultural context, while minimizing the occurrence of technical issues that could interfere with the testing sessions. Improvements to the eTIMSS Online Translation System began during countries' preparations for the eTIMSS Pilot / Item Equivalence Study and continued in advance of the Field Test. As described earlier, the eTIMSS Online Translation System was a source of frustration for some countries in conducting the eTIMSS Pilot / Item Equivalence Study. Generally, countries found the translation process to be very time consuming due to limitations in the ability to implement national adaptations. Some countries also expressed frustration with the user interface of the Translation System.

In collaboration with the IEA, the TIMSS & PIRLS International Study Center established processes for the TIMSS 2019 Field Test to ease the translation process for countries, particularly for countries with right-to-left formatted languages. To prepare instruments for the Field Test, countries could opt to receive an international Arabic source version of all instruments, and countries were be able to horizontally flip individual images and equations as desired. In addition, the number pad was fully adaptable for each country's context. IEA Hamburg improved the interface of the eTIMSS Translation System to ensure the system is easy to use by translators and could allow for variations in country languages and unique characters.

Improving the eTIMSS Player

IEA Hamburg made many updates to the eTIMSS Player to improve the reliability of the software, and spent great effort testing the software components before releasing to countries for the Field Test. In particular, staff worked to ensure items were technically accessible for all types of responses, regardless of device, language, and national context. The TIMSS & PIRLS International Study Center also made enhancements to the eTIMSS user interface to improve the experience for students. To further minimize the potential for student confusion or difficulty in navigating eTIMSS or responding to items, the directions module was improved and expanded.

Feedback about the eTIMSS Player from countries participating in the eTIMSS Pilot / Item Equivalence Study mostly applied to tablet devices, including the ability of students to exit the eTIMSS Player during testing and the need for programming the on-screen keyboard equipped by the device. Additional feedback about the eTIMSS Player general to both PC and tablet devices concerned the functioning of the response inputs for each item, the means for navigation within the eTIMSS interface, and the data upload procedures.

Some NRCs reported that text fields for constructed response items were too small for typing a full response, particularly for character-based languages. In some instances, students could not read their entire typed response with a very short answer field. For newly developed items for TIMSS 2019, the TIMSS & PIRLS International Study Center made efforts to anticipate the length of responses and provide larger answer fields where necessary. IEA Hamburg made additional technical adjustments for countries with character-based languages that may require larger text fields.

In response to difficulties encountered during the eTIMSS Pilot / Item Equivalence Study in uploading data, IEA Hamburg also improved the eTIMSS Player for the Field Test to indicate when data upload was successful. A majority of countries expressed desire to have feedback given on the device when uploading student response data to the IEA server. Some countries reported issues of missing records or item responses after upload, which was resolved with a second data upload.

Many countries provided feedback on the means for navigation in the eTIMSS interface. For example, the left side progress bar has sequentially numbered buttons for each item screen. At the start of the assessment, all buttons are green. Once a student has visited a screen, the button turns dark, regardless of whether the student interacted with the item, or simply left it blank. In order for students to keep better track of which items they have completed, NRCs recommended having a different color for visited screens with no answers provided. This would better resemble the experience on paper, where students can mark items to skip and return later.

Writing Manuals

In preparation for the TIMSS 2019 Field Test, staff at the TIMSS & PIRLS International Study Center wrote detailed manuals and supplemental reference materials for National Research Coordinators (NRCs), School Coordinators, and Test Administrators to anticipate and efficiently solve problems while conducting the Field Test. The experiences of countries in conducting the eTIMSS Pilot / Item Equivalence Study provided information to the TIMSS & PIRLS International Study Center about the various problems that could occur in distributing, setting up, and administering eTIMSS.

These types of issues reported emphasized the importance of advance planning for the assessment and the need to provide detailed instructions for School Coordinators and Test Administrators to troubleshoot issues that may occur before or during the test session. The TIMSS & PIRLS International Study Center wrote Field Test manuals with more detailed procedures for distributing the assessment to schools and for setting up a larger variety of devices. The manuals also included instructions for solving a wider range of potential issues. In addition, a new "server method" was introduced to allow countries to distribute the eTIMSS software within schools using a local area network (LAN), rather than installing on each individual computer.

In addition to reporting issues, some National Research Coordinators provided explicit suggestions to include in manuals for the Field Test. For example, one NRC suggested reminding School Coordinators and Test Administrators to charge portable devices between testing sessions. Some countries also noted that device settings should be managed before testing, including turning off device sound and disabling any "autocorrect" or "autocomplete" keyboard features that may suggest answers for students. In addition, some languages require the use of multiple keyboards that allow for mathematical notation. For example, some Arabic-speaking countries use "x" and "y" letters for mathematical notation, so needed both Arabic and English keyboards active for testing.

It should be noted that many of the issues reported about distributing and setting-up devices in schools for the eTIMSS Pilot / Item Equivalence Study could not clearly be attributed to limitations of the eTIMSS Player or to the SystemCheck program. Therefore, the TIMSS & PIRLS International Study Center and IEA Hamburg made

additional efforts to test all software components on various types of devices before releasing the software to countries for the Field Test. The eAssessment developers at IEA Hamburg hypothesized that many of the ambiguous issues reported by countries were related to the devices used by schools or the IT network set up at schools. For example, one NRC reported that the SystemCheck program indicated a PC met all eTIMSS requirements, but when run again, gave a different diagnosis related to screen resolution or available memory. Another country reported a few instances of devices crashes during testing, despite the improvements made to the eAssessment system components prior to release for the eTIMSS Pilot / Item Equivalence Study. Programmers at IEA Hamburg hypothesized that such issues may have occurred because the system software of the devices was not up-to-date before installing the eTIMSS software.

Developing Digitally Enhanced Items

Item development for TIMSS 2019 involved developing pools of both paper items (for paperTIMSS countries) and digital items (for eTIMSS countries) that have approximately the same coverage of the *TIMSS 2019 Assessment Frameworks* (Mullis & Martin, 2017). This process began by first developing sets of digitally enhanced items in mathematics and science that capitalize upon technology to improve the assessment experience for students. New item types for the Field Test include drag and drop, selectable, sortable, and drop-down menu inputs for responding to items.

In addition, certain item features deemed unsuccessful for the eTIMSS Pilot / Item Equivalence Study were eliminated. For example, canvas item types, where students draw or write responses with a mouse, finger, or stylus, depending on the device, were at times difficult for students to respond and often unable to be scored using the IEA's

Scoring System. Similar to the experiences of students during the AIR Cognitive Labs and the eTIMSS prePilot, the results of the Pilot / Item Equivalence Study indicated that students became frustrated with drawing and writing and were unable to provide the type of response that these items elicited. For example, completing a bar graph with labels was very difficult for students. Many students left these items blank. In light of these issues and the lack of ability to score the items accurately in the IEA Scoring System, the TIMSS & PIRLS International Study Center removed the canvas item type for the TIMSS 2019 Field Test in lieu of developing a line tool. The line tool will provide students similar capabilities to the draw tool and is expected to be easier for students to use.

As discussed earlier in this chapter, reducing scrolling on trend items was a major goal in converting the trend items to a digital format as well as in developing new items for TIMSS 2019. Despite efforts to reduce mode effects related to scrolling after the eTIMSS prePilot, the trend items that required substantial scrolling (most often two-page items on paper) were found to be unsuccessful for the eTIMSS Pilot / Item Equivalence Study. Even for multiple-choice items, relatively high omission rates on these items for eTIMSS compared to paperTIMSS suggest that students did not see entire parts of an item and so left them blank.

Unfortunately, eliminating scrolling completely is not feasible, particularly for trend items. In addition to reducing the amount of text and size of graphics to reduce scrolling, item development for TIMSS 2019 included additional efforts to indicate to students that there is more content on a screen. Some items in PSI tasks are programmed to alert students with a pop-up message when part or all of an item has not been activated

(e.g., “Are you sure you are finished?”). Other solutions are also under consideration for future data collection efforts, including a multiscreen feature for multi-part items.

Chapter 3: Methodology and Results of the eTIMSS Pilot / Item Equivalence Study

Overview

The experimental research design for the eTIMSS Pilot / Item Equivalence Study produced student achievement estimates based on the trend items from TIMSS 2015 at the fourth grade and eighth grade in two modes of assessment delivery—paperTIMSS and eTIMSS. Each student sampled for the eTIMSS Pilot / Item Equivalence Study received one paperTIMSS booklet as well as one digital item block combination for eTIMSS. Following administration and scoring of the instruments, three phases of analysis were conducted to examine the effect of the digital mode of administration on TIMSS trend items and scale scores:

1. A priori analysis of item equivalence
2. Item analysis based on classical item statistics
3. Scale score analysis of TIMSS IRT estimates

The author of this dissertation conducted Phase 1 and helped conduct Phase 2 with a team at the TIMSS & PIRLS International Study Center. Staff from Educational Testing Service (ETS) verified the results of the item analysis and also produced the proficiency score estimates and conducted the scale score analysis in Phase 3. The dissertation author replicated and extended the results of Phase 3.

After describing the sample design and research design for the eTIMSS Pilot / Item Equivalence Study, this chapter describes the procedures followed for each of the

three phases of analysis conducted and the results. Results are summarized at the end of the chapter.

Sample Design

Twenty-five countries participated in the eTIMSS Pilot / Item Equivalence Study, with 24 countries participating at the fourth grade and 13 countries at the eighth grade (see Exhibit 3.1). Each country selected a purposive sample of 800 students at each grade that included students with a range of abilities and backgrounds. Chile participated informally at the fourth grade with a small sample of students.

Exhibit 3.1: Countries Participating in the eTIMSS Pilot / Item Equivalence Study

Bulgaria (4)	Japan (4 and 8)	Qatar (4 and 8)
Chile (4)	Korea, Rep. of (4 and 8)	Russian Federation (4 and 8)
Croatia (4)	Lithuania (4 and 8)	Spain (4)
Czech Republic (4)	Malaysia (8)	Sweden (4 and 8)
Denmark (4)	Morocco (4 and 8)	Turkey (4)
Finland (4 and 8)	Netherlands (4)	United Arab Emirates (4 and 8)
France (4)	Norway (4)	United States (4 and 8)
Germany (4)	Oman (4 and 8)	
Italy (4 and 8)	Portugal (4)	

Grade(s) of participation appear in parentheses.
Chile participated informally with a small sample of students at the fourth grade.

Instruments

Achievement Booklets / Block Combinations

The complete set of trend items from TIMSS 2015 were administered for the eTIMSS Pilot / Item Equivalence Study—187 items at the fourth grade and 232 items at

the eighth grade. For each grade and subject, three of the eight secured trend blocks were introduced in TIMSS 2011 and the other five were introduced in TIMSS 2015. These trend items will be the basis for the link from TIMSS 2015 to TIMSS 2019.

The item blocks were developed for TIMSS 2011 and TIMSS 2015 with approximately 10 to 14 items per block at the fourth grade and 12 to 18 items per block at the eighth grade, and according to the assessment framework targets for that cycle (Mullis & Martin, 2013; Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009). Each item block mimicked the distribution of item types as well as the content and cognitive skills that the entire assessment is meant to cover. Fourth grade students were expected to spend 18 minutes on each block and eighth grade students were expected to spend 22 ½ minutes per block.

For each TIMSS assessment cycle, the entire item pool is developed so that approximately half the score points come from multiple-choice items and the other half from constructed response items. Multiple-choice items are each worth 1 score point, with some compound multiple-choice items—or multi-part multiple-choice items—worth 2 score points. In accordance with the guidelines for item development each cycle, constructed response items were worth 1 or 2 score points depending on the complexity of the item.

As discussed in Chapter 2, the TIMSS & PIRLS International Study Center converted the trend items to the digital eTIMSS format so that students could respond in similar ways to that on paper. Converting multiple-choice items was relatively simple, with the use of buttons for answer options that students selected with a mouse, finger, or stylus, depending on the device.

Digital versions of constructed response items administered for the eTIMSS Pilot / Item Equivalence Study have different "input types," which specify the response action required of students to answer the item in their digital format. On paper, constructed response trend items require students to use a pencil to write written responses, show work, draw graphs and diagrams, and write mathematical equations. The eTIMSS Pilot / Item Equivalence Study included the following digital item types according to their constructed response inputs:

- **Keyboard**—students use the full keyboard equipped by the delivery device (either external or on-screen) to type responses, including mathematical equations.
- **Number pad**—students use an on-screen number keypad to enter a numerical response. For the eTIMSS Pilot / Item Equivalence Study, the number pad included the digits 0 through 9, as well as a decimal point (.), negative sign (-), and a division symbol for fractions (/).
- **Canvas**—students use mouse, finger, or stylus, depending on the device and their preference, to show work, draw, or label diagrams

Although students found the "canvas" item types to be difficult in the eTIMSS prePilot, this feature was irreplaceable for the trend items where students needed to draw graphs or diagrams or show work. Without this feature, some items would have to undergo substantial alterations that could threaten the equivalence of the item between paper and digital modes of administration. Although data were lost for most of these items due to limitations of the IEA Scoring System, some canvas items did not require scorers to see the full item screen and received scores.

Exhibit 3.2 presents the number and percentage of each digital item type administered for the eTIMSS Pilot / Item Equivalence Study by grade and subject. The counts include the canvas items, most of which could not be scored due to the issues discussed previously. Some items used a combination of input types. These items are reported in Exhibit 3.2 according to the more prominent input used to respond (see footnote below exhibit).

Exhibit 3.2: Distribution of Items in the eTIMSS Pilot / Item Equivalence Study by Digital Item Type

Digital Item Type	Mathematics		Science	
	Count	Percent of Items	Count	Percent of Items
Fourth Grade				
Multiple-Choice	42	46%	47	49%
Compound Multiple-Choice	2	2%	5	5%
Keyboard	6	7%	42	44%
Number Pad	30	33%	0	0%
Canvas	12	13%	1	1%
Total	92	100%	95	100%
Eighth Grade				
Multiple-Choice	62	54%	59	50%
Compound Multiple-Choice	0	0%	7	6%
Keyboard	12	11%	48	41%
Number Pad	33	29%	2	2%
Canvas	7	6%	2	2%
Total	114	100%	118	100%

Because of rounding some results may appear inconsistent.
At the fourth grade, 1 mathematics item classified as "canvas" includes both keyboard and canvas inputs.
At the eighth grade, 1 mathematics item classified as "number pad" includes both keyboard and number pad inputs and 2 mathematics items classified as "canvas" include both keyboard and canvas inputs.

Student Questionnaire

Following the delivery of the digital item block combinations, fourth and eighth grade students received a brief student questionnaire through the eTIMSS Player. The questionnaire included items asking about:

- Students' gender
- Number of books in their home
- Parents' highest level of education (eighth grade only)
- Access to a computer or tablet and Internet at home
- Access to a computer or tablet and Internet at school
- Time spent using a computer or tablet each day
- Frequency that students use a computer or tablet to do five school-related activities
- Students' agreement with three statements about their ability to use technology
- Students' confidence to perform five tasks on a digital device

Chapter 4 includes full item descriptions for variables of interest to the dissertation.

Research Design

Exhibit 3.3 shows the trend item blocks assigned to each paper booklet and its equivalent eTIMSS block combination in digital format. Consistent with the rotated counterbalanced design used for each TIMSS cycle, each item block appears in two booklets and block combinations in different positions to control for position effects and to provide a mechanism for linking student responses across test forms for scaling (Martin, Mullis, & Foy, 2013). Each booklet and block combination is divided into two

parts and contains two blocks of mathematics items (beginning with "M") and two blocks of science items (beginning with "S"). Half begin with two blocks of mathematics items and half begin with two blocks of science items.

Using the same eight-booklet design allowed for closely replicating the presentation of the trend blocks to the TIMSS 2015 assessment. Six of the eight booklets are identical to booklets administered in 2015. The other two booklets contain two trend item blocks that were also paired together in 2015, but with two other blocks (M04 and S04) added to the booklets to replace blocks that were removed after the 2015 cycle.

Exhibit 3.3: eTIMSS Pilot / Item Equivalence Study Booklet/Block Combination Design—Fourth and Eighth Grades

Paper Booklet	Digital Block Combination	Trend Item Blocks			
		Part 1		Part 2	
Booklet 1	ET19PTBC01	M04	M08	S04	S08
Booklet 2*	ET19PTBC02	S08	S09	M08	M09
Booklet 3*	ET19PTBC03	M09	M10	S09	S10
Booklet 4*	ET19PTBC04	S10	S11	M10	M11
Booklet 5*	ET19PTBC05	M11	M12	S11	S12
Booklet 6*	ET19PTBC06	S12	S13	M12	M13
Booklet 7*	ET19PTBC07	M13	M14	S13	S14
Booklet 8	ET19PTBC08	S14	S04	M14	M04

* Booklet identical to booklet administered for TIMSS 2015.

Exhibit 3.4 shows the counterbalanced rotation scheme for eTIMSS Pilot / Item Equivalence Study data collection, such that each student sampled received one test booklet in the traditional paper-and-pencil format (paperTIMSS) and one digital item block combination delivered through the eTIMSS Player on a PC or tablet device (eTIMSS). In each country, half of students received paperTIMSS first (Booklets 1–8),

and half of students received eTIMSS first (ET19PTBC01–08). The two test sessions occurred either within the same day or across two consecutive days.

Exhibit 3.4: eTIMSS Pilot / Item Equivalence Study Booklet/Block Combination Rotation Scheme—Fourth and Eighth Grades

Rotation	Session 1	Session 2
1	Booklet 1	ET19PTBC03
2	ET19PTBC04	Booklet 2
3	Booklet 3	ET19PTBC05
4	ET19PTBC06	Booklet 4
5	Booklet 5	ET19PTBC07
6	ET19PTBC08	Booklet 6
7	Booklet 7	ET19PTBC01
8	ET19PTBC02	Booklet 8

Phase 1: A Priori Analysis of Item Equivalence

The eTIMSS Pilot / Item Equivalence Study included a qualitative a priori analysis of item equivalence. The analysis involved developing a set of explicit criteria for classifying the items according to their differences across paper and digital formats. The dissertation author and two additional staff from the TIMSS & PIRLS International Study Center classified the trend items according to their hypothesized likelihood for being "strongly equivalent" or "invariant" between paper and digital delivery modes.

A Priori Analysis Procedure

Before the analysis, the author defined preliminary item classification descriptions based on the results of the AIR Cognitive Labs and the eTIMSS prePilot, as well as mode effect literature relevant to the types of items in the eTIMSS Pilot / Item Equivalence

Study instruments. The following types of items or features of items were of particular interest in the analysis:

- Differences in presentation between paper and digital formats (Pommerich, 2004), including items that required significant changes to the formatting to render on a digital interface (Sandene et al., 2005)
- Complex graphs or diagrams (Mazzeo & Harvey, 1988) or heavy reading possibly requiring greater cognitive processing (Chen, Cheng, Chang, Zheng, & Huang, 2014; Noyes & Garland, 2008)
- Scrolling required to view all parts of the item (Bridgeman et al., 2003; Pommerich, 2004; Way et al., 2016), particularly for science items when the student must refer to earlier parts of the item to formulate a response (Pommerich, 2004)
- Constructed response items requiring long explanations (Strain-Seymour et al., 2013), due to differences in students' typing abilities (Russell, 1999), typing fatigue that could occur with an on-screen keyboard (Pisacreta, 2013), or the potential for scoring bias between paperTIMSS and eTIMSS item responses (Horkay et al., 2006; Russell, 2002)
- Constructed response items requiring calculations by hand or with a calculator, which may require students to transcribe calculations from scratch paper to the PC or tablet to receive full credit (Johnson & Green, 2006)
- Items with numerical answers requiring the number pad to input the response. This was of particular concern for answers involving decimals, four-digit numbers, negative numbers, and fractions, due to limitations of the eTIMSS

Translation System at the time of the study. At the fourth grade in particular, inputting complex numbers such as fractions was thought to be cumbersome for students.

- Difficult response modes, including items requiring students to draw, label, or manipulate features (Sandene et al., 2005; Strain-Seymour et al., 2013)

The author established separate criteria for the fourth and eighth grades based on the assumption that eighth grade students have more experience with using PCs and tablets. Eighth grade items also tend to involve more reading and calculation to solve problems, and often require longer typed or written answers.

Two raters examined the international version of each trend item in paper, tablet, and PC formats, along with scoring guides for constructed response items to understand what is required for a correct response. The raters refined the pre-developed criteria into detailed descriptions and used them to classify each item into one of four types described in Exhibit 3.5—"Identical," "Nearly Identical," "Worrisome," or "Severe." When the two raters disagreed, a third rater who was also familiar with the trend items made the final classification.

During the analysis, the author refined and expanded the criteria for particular nuances of the items. For example, some items had directional language that may not apply to the digital format, such as "mark an X" when the digital item required the X's to be typed.

Exhibit 3.5: A Priori Trend Item Classifications for the eTIMSS Pilot / Item Equivalence Study

Classification	Description of Criteria	
	Fourth Grade	**Eighth Grade**
Identical	• Looks identical across paper and digital formats • Negligible adjustments made to graphics or layout for digital format • Multiple-choice items with no graphics or with small graphics • Constructed response items requiring 1–2 word responses	• Looks identical across paper and digital formats • Negligible adjustments made to graphics or layout for digital format • Multiple-choice items with no graphics or with small graphics • Constructed response items requiring 1–3 word responses
Nearly Identical	• Minor adjustments made to graphics or layout to fit screen • Compound multiple-choice items (minor layout differences) • Constructed response items requiring short responses (approximately 3 words)	• Minor adjustments made to graphics or layout to fit screen • Compound multiple-choice items (minor layout differences) • Heavy reading requirement • Constructed response items requiring short responses (approximately 4 words) • Number pad response mode for eTIMSS (except fraction answers)
Worrisome	• Some scrolling required • Moderate changes made to layout to fit screen • Heavy reading requirement • Constructed response items requiring long responses or equation answers • Items requiring calculation or transcription from scratch paper • Number pad response mode for eTIMSS (except fraction answers) • Directional language that may not apply to digital mode	• Some scrolling required • Moderate changes made to layout to fit screen • Constructed response items requiring long responses or equation answers • Items requiring calculation or transcription from scratch paper • Number pad response mode for eTIMSS with fraction answers • Canvas response mode for eTIMSS requiring students to make simple marks with draw tool • Directional language that may not apply to digital mode
Severe	• Severe scrolling required • Substantial changes made to layout to fit screen • Number pad response mode for eTIMSS with fraction answers • Canvas response mode for eTIMSS requiring students to draw or label graphs with the draw tool	• Severe scrolling required • Substantial changes made to layout to fit screen • Canvas response mode for eTIMSS requiring students to draw or label graphs with the draw tool

A Priori Analysis Results

Exhibit 3.6 displays the results of the a priori analysis. Items classified as "Identical" were hypothesized to not exhibit any potential mode effects and have approximately the same measurement properties between paperTIMSS and eTIMSS. Items classified as "Nearly Identical" or "Worrisome" were hypothesized to *possibly* show mode effects, with "Worrisome" items being more likely to show differences in behavior. Items classified as "Severe" were hypothesized to exhibit mode differences affecting measurement equivalence.

Exhibit 3.6: Distribution of Items in the eTIMSS Pilot / Item Equivalence Study by A Priori Classification

Classification	Mathematics		Science	
	Count	Percent of Items	Count	Percent of Items
Fourth Grade				
Identical	26	28%	36	38%
Nearly Identical	17	18%	38	40%
Worrisome	33	36%	11	12%
Severe	16	17%	10	11%
Total	92	100%	95	100%
Eighth Grade				
Identical	46	40%	53	45%
Nearly Identical	33	29%	44	37%
Worrisome	27	24%	11	9%
Severe	8	7%	10	8%
Total	114	100%	118	100%

Because of rounding some results may appear inconsistent.

Overall, the results of the a priori analysis indicate that the majority of the trend items at each grade were "Identical" and could be expected to perform the same for both paperTIMSS and eTIMSS. These results provided face validity evidence that the items measure the same constructs in both modes, and that efforts to keep items the same were

mostly successful. Fewer items were hypothesized to show mode effects at the eighth grade (7% for mathematics and 8% for science) compared to the fourth grade (17% and 11% for mathematics and science, respectively). At the eighth grade, 40 percent of mathematics items and 45 percent of science items were "Identical" and predicted to be equivalent for paperTIMSS and eTIMSS, compared to 28 percent of mathematics items and 38 percent of science items at the fourth grade. Across the two grades, fewer science items were "Worrisome" or "Severe" and hypothesized to show mode effects compared to mathematics items.

Staff at the TIMSS & PIRLS International Study Center further refined the a priori classifications for the item analysis described in the next section.

Phase 2: Item Analysis

The TIMSS & PIRLS International Study Center conducted an analysis of classical item statistics to examine the measurement equivalence of the TIMSS 2019 trend items between paper and digital modes of administration. A preliminary review of the item statistics also identified any items that were not well suited for the eTIMSS platform, due to technical issues or the inability of students to provide the type of response that the item was meant to elicit on paper.

The author was part of the team from the TIMSS & PIRLS International Study Center to produce the item statistics and review the results, and staff from ETS replicated the item statistics. The procedure included producing three sets of item statistics for each trend item based on paperTIMSS responses and eTIMSS responses, respectively, as well as the difference between modes for each statistic. The author computed average item

statistics for groups of trend items by grade and subject, as well as by digital item or input type. The following statistics were of interest during the analysis:

- percent correct (item difficulty)
- point-biserial correlations (item discrimination)
- missing rates (percent omitted, percent not reached)

Item Analysis Procedure

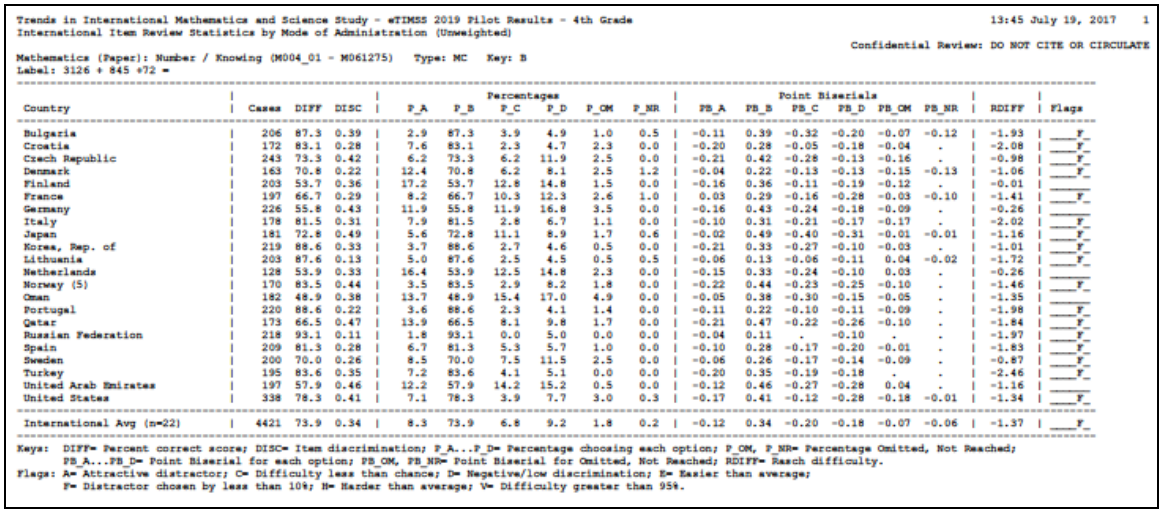
Reviewing the Item Statistics

The TIMSS & PIRLS International Study Center produced item statistics by country—called "item almanacs"—for all items included in the eTIMSS Pilot / Item Equivalence Study, based on paperTIMSS and eTIMSS response data, respectively. Two sets of item statistics were produced for 187 fourth grade items (92 mathematics items and 95 science items) and 232 eighth grade items (114 mathematics items and 118 science items). Any country-specific recodes made to items for the TIMSS 2015 assessment were also made to the eTIMSS Pilot / Item Equivalence Study data, including item deletions or constructed response items with score category recodes (see Foy, Martin, Mullis, Yin, Centurino, & Reynolds, 2016 for details).

Exhibits 3.7 and 3.8 show examples of the statistics produced for a multiple-choice item based on paperTIMSS data and a constructed response item based on eTIMSS data, respectively. Each item almanac page includes statistics for each country and the international average with each country weighted equally. The mode of administration is specified in parentheses in the top left corner of each table, following the subject of the item—mathematics or science. For example, Exhibit 3.7 shows statistics for a paperTIMSS mathematics item, indicated by "Mathematics (Paper)."

Exhibit 3.8 shows statistics for an eTIMSS science item, indicated by "Science (Electronic)." Other information about the item is included above the table, including the item format (multiple-choice or constructed response), maximum score points (1 or 2 score points, depending on the item), and the scoring key for multiple-choice items (A, B, C, or D).

Exhibit 3.7: Example Item Statistics for a Multiple-Choice Item—paperTIMSS



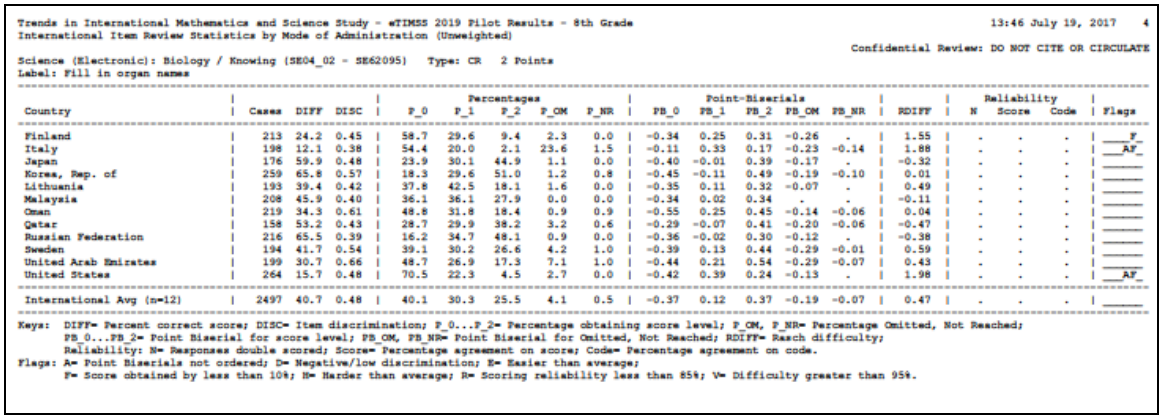
Trends in International Mathematics and Science Study - eTIMSS 2019 Pilot Results - 4th Grade — 13:45 July 19, 2017 — 1
International Item Review Statistics by Mode of Administration (Unweighted)
Confidential Review: DO NOT CITE OR CIRCULATE

Mathematics (Paper): Number / Knowing (M004_01 - M061275) Type: MC Key: B
Label: 3126 + 845 +72 =

Country	Cases	DIFF	DISC	Percentages P_A	P_B	P_C	P_D	P_OM	P_NR	Point Biserials PB_A	PB_B	PB_C	PB_D	PB_OM	PB_NR	RDIFF	Flags
Bulgaria	206	87.3	0.39	2.9	87.3	3.9	4.9	1.0	0.5	-0.11	0.39	-0.32	-0.20	-0.07	-0.12	-1.93	____F_
Croatia	172	83.1	0.28	7.6	83.1	2.3	4.7	2.3	0.0	-0.20	0.28	-0.05	-0.18	-0.04	.	-2.08	____F_
Czech Republic	243	73.3	0.42	6.2	73.3	6.2	11.9	2.5	0.0	-0.21	0.42	-0.28	-0.13	-0.16	.	-0.98	____F_
Denmark	163	70.8	0.22	12.4	70.8	6.2	8.1	2.5	1.2	-0.04	0.22	-0.13	-0.13	-0.15	-0.13	-1.06	____F_
Finland	203	53.7	0.36	17.2	53.7	12.8	14.8	1.5	0.0	-0.16	0.36	-0.11	-0.19	-0.12	.	-0.01	_____
France	197	66.7	0.29	8.2	66.7	10.3	12.3	2.6	1.0	0.03	0.29	-0.16	-0.28	-0.03	-0.10	-1.41	____F_
Germany	226	55.8	0.43	11.9	55.8	11.9	16.8	3.5	0.0	-0.16	0.43	-0.24	-0.18	-0.09	.	-0.26	_____
Italy	178	81.5	0.31	7.9	81.5	2.8	6.7	1.1	0.0	-0.10	0.31	-0.21	-0.17	-0.17	.	-2.02	____F_
Japan	181	72.8	0.49	5.6	72.8	11.1	8.9	1.7	0.6	-0.02	0.49	-0.40	-0.31	-0.01	-0.01	-1.16	____F_
Korea, Rep. of	219	88.6	0.33	3.7	88.6	2.7	4.6	0.5	0.0	-0.21	0.33	-0.27	-0.10	-0.03	.	-1.01	____F_
Lithuania	203	87.6	0.13	5.0	87.6	2.5	4.5	0.5	0.5	-0.06	0.13	-0.06	-0.11	0.04	-0.02	-1.72	____F_
Netherlands	128	53.9	0.33	16.4	53.9	12.5	14.8	2.3	0.0	-0.15	0.33	-0.24	-0.10	0.03	.	-0.26	_____
Norway (5)	170	83.5	0.44	3.5	83.5	2.9	8.2	1.8	0.0	-0.22	0.44	-0.23	-0.25	-0.10	.	-1.46	_____
Oman	182	48.9	0.38	13.7	48.9	15.4	17.0	4.9	0.0	-0.05	0.38	-0.30	-0.15	-0.05	.	-1.35	_____
Portugal	220	88.6	0.22	3.6	88.6	2.3	4.1	1.4	0.0	-0.11	0.22	-0.10	-0.11	-0.09	.	-1.98	____F_
Qatar	173	66.5	0.47	13.9	66.5	8.1	9.8	1.7	0.0	-0.21	0.47	-0.22	-0.26	-0.10	.	-1.84	____F_
Russian Federation	218	93.1	0.11	1.8	93.1	0.0	5.0	0.0	0.0	-0.04	0.11	.	-0.10	.	.	-1.97	____F_
Spain	209	81.3	0.28	6.7	81.3	5.3	5.7	1.0	0.0	-0.10	0.28	-0.17	-0.20	-0.01	.	-1.83	____F_
Sweden	200	70.0	0.26	8.5	70.0	7.5	11.5	2.5	0.0	-0.06	0.26	-0.17	-0.14	-0.09	.	-0.87	____F_
Turkey	195	83.6	0.35	7.2	83.6	4.1	5.1	0.0	0.0	-0.20	0.35	-0.19	-0.18	.	.	-2.46	____F_
United Arab Emirates	197	57.9	0.46	12.2	57.9	14.2	15.2	0.5	0.0	-0.12	0.46	-0.27	-0.28	0.04	.	-1.16	_____
United States	338	78.3	0.41	7.1	78.3	3.9	7.7	3.0	0.3	-0.17	0.41	-0.12	-0.28	-0.18	-0.01	-1.34	____F_
International Avg (n=22)	4421	73.9	0.34	8.3	73.9	6.8	9.2	1.8	0.2	-0.12	0.34	-0.20	-0.18	-0.07	-0.06	-1.37	____F_

Keys: DIFF= Percent correct score; DISC= Item discrimination; P_A...P_D= Percentage choosing each option; P_OM, P_NR= Percentage Omitted, Not Reached;
PB_A...PB_D= Point Biserial for each option; PB_OM, PB_NR= Point Biserial for Omitted, Not Reached; RDIFF= Rasch difficulty.
Flags: A= Attractive distractor; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
F= Distractor chosen by less than 10%; H= Harder than average; V= Difficulty greater than 95%.

Exhibit 3.8: Example Item Statistics for a Constructed Response Item—eTIMSS



Trends in International Mathematics and Science Study - eTIMSS 2019 Pilot Results - 8th Grade — 13:46 July 19, 2017 — 4
International Item Review Statistics by Mode of Administration (Unweighted)
Confidential Review: DO NOT CITE OR CIRCULATE

Science (Electronic): Biology / Knowing (SE04_02 - SE62095) Type: CR 2 Points
Label: Fill in organ names

Country	Cases	DIFF	DISC	Percentages P_0	P_1	P_2	P_OM	P_NR	Point-Biserials PB_0	PB_1	PB_2	PB_OM	PB_NR	RDIFF	Reliability N	Score	Code	Flags
Finland	213	24.2	0.45	58.7	29.6	9.4	2.3	0.0	-0.34	0.25	0.31	-0.26	.	1.55	.	.	.	____F_
Italy	198	12.1	0.38	54.4	20.0	2.1	23.6	1.5	-0.11	0.33	0.17	-0.23	-0.14	1.88	.	.	.	__AF_
Japan	176	59.9	0.48	23.9	30.1	44.9	1.1	0.0	-0.40	-0.01	0.39	-0.17	.	-0.32	.	.	.	_____
Korea, Rep. of	259	65.8	0.57	18.3	29.6	51.0	1.2	0.8	-0.45	-0.11	0.49	-0.19	-0.10	0.01	.	.	.	_____
Lithuania	193	39.4	0.42	37.8	42.5	18.1	1.6	0.0	-0.35	0.11	0.32	-0.07	.	0.49	.	.	.	_____
Malaysia	208	45.9	0.40	36.1	36.1	27.9	0.0	0.0	-0.34	0.02	0.34	.	.	-0.11	.	.	.	_____
Oman	219	34.3	0.61	48.8	31.8	18.4	0.9	0.9	-0.55	0.25	0.45	-0.14	-0.06	0.04	.	.	.	_____
Qatar	158	53.2	0.43	28.7	29.9	38.2	3.2	0.6	-0.29	-0.07	0.41	-0.20	-0.06	-0.47	.	.	.	_____
Russian Federation	216	65.5	0.39	16.2	34.7	48.1	0.9	0.0	-0.36	-0.02	0.30	-0.12	.	-0.38	.	.	.	_____
Sweden	194	41.7	0.54	39.1	30.2	26.6	4.2	1.0	-0.39	0.13	0.44	-0.29	-0.01	0.59	.	.	.	_____
United Arab Emirates	199	30.7	0.66	48.7	26.9	17.3	7.1	1.0	-0.44	0.21	0.54	-0.29	-0.07	0.43	.	.	.	_____
United States	264	15.7	0.48	70.5	22.3	4.5	2.7	0.0	-0.42	0.39	0.24	-0.13	.	1.98	.	.	.	__AF_
International Avg (n=12)	2497	40.7	0.48	40.1	30.3	25.5	4.1	0.5	-0.37	0.12	0.37	-0.19	-0.07	0.47	.	.	.	_____

Keys: DIFF= Percent correct score; DISC= Item discrimination; P_0...P_2= Percentage obtaining score level; P_OM, P_NR= Percentage Omitted, Not Reached;
PB_0...PB_2= Point Biserial for score level; PB_OM, PB_NR= Point Biserial for Omitted, Not Reached; RDIFF= Rasch difficulty;
Reliability: N= Responses double scored; Score= Percentage agreement on score; Code= Percentage agreement on code.
Flags: A= Point Biserials not ordered; D= Negative/low discrimination; E= Easier than average;
F= Score obtained by less than 10%; H= Harder than average; R= Scoring reliability less than 85%; V= Difficulty greater than 95%.

Each almanac page reports the number of respondents in each country (Cases), the item difficulty (DIFF), and item discrimination (DISC). For multiple-choice items,

the percentage of students choosing each answer option is reported (P_A, P_B, P_C, P_D), and for constructed response items, the percentage of students at each score level (0, 1, or 2 score points, depending on the item) is reported (P_0, P_1, P_2). The tables also include the percentage of students who omitted the item (P_OM) and the percentage of students who did not reach the item (P_NR). An item response was considered to be "not reached" for a case if the student did not answer the item before it and did not answer any subsequent items within the part of the assessment booklet or block combination (Foy et al., 2016). Item flags (Flags) indicate items with unusual psychometric properties that may suggest a problem with the item. The item almanacs also included an estimate of item difficulty based on a Rasch one-parameter IRT model (RDIFF) for each country, as well as point-biserial correlations for each response category.

The author used the statistics from the paperTIMSS and eTIMSS item almanacs to compute "difference statistics" for each item. The difference between paper and digital modes for each of the statistics were computed by subtracting the eTIMSS statistic from the paperTIMSS statistic (e.g., $DIFF_{paper} - DIFF_{eTIMSS}$). This produced classical measures of mode effect statistics for each item by country. Difference statistics for item discrimination, percent omitted, and percent not reached were calculated the same way.

The author created almanacs of item difference statistics that closely resembled the paperTIMSS and eTIMSS item almanacs. Exhibit 3.9 shows an example item difference almanac page produced for a compound multiple-choice item. The almanac reports most of the same item statistics as are in Exhibits 3.7 and 3.8—item difficulty, item discrimination, percentage of students by response category, percent omitted, and

percent not reached. Each page includes the number of student responses to the item for both paperTIMSS and eTIMSS. Positive differences indicate that the item statistic was greater for paperTIMSS than for eTIMSS, and negative values indicate that the item statistic was greater for eTIMSS than for paperTIMSS.

Above each item difference table, the digital item type—or the type of input for responding to the item—is specified, as well as the a priori item classification. The far right "FLAG" column indicates whether there was a deviation between paper and digital versions of the item reported by a country representative or by a staff member at IEA Hamburg. The author compiled the item-specific feedback from each country to reference during item review.

Exhibit 3.9: Example Difference Statistics for a Compound Multiple-Choice Item

eTIMSS 2019 Pilot Results - 8th Grade — ITEM EQUIVALENCE — 122
International Item Review Statistics — (paperTIMSS - eTIMSS) — Confidential Review: DO NOT CITE OR CIRCULATE

Block Seq:	Unique ID:		Format:	Key:	MaxPts:	Input Type:	Classification: Nearly Identical
S004_08	S062018	SE62018	CR	X	2	Compound Multiple-Choice	

					Percentages							
Country	n paperTIMSS	n eTIMSS	DIFF	DISC	0 / A	1 / B	2 / C	D	OMIT	NOT REACHED	RDIFF	FLAG
Finland	212	213	-1.8	0.1	4.1	-2.8	-0.4		-0.9	0.0	-0.12	-
Italy	183	198	0.9	-0.0	-3.4	2.4	-0.3		1.3	-1.0	-0.17	-
Japan	166	176	2.4	0.2	-2.1	2.7	1.0		-1.6	0.0	-0.15	D
Korea, Rep. of	250	259	-1.2	0.1	2.2	-4.3	0.9		1.2	-0.8	-0.27	-
Lithuania	185	193	2.2	0.0	-4.0	5.7	-0.7		-1.0	0.0	-0.16	-
Malaysia	220	208	-0.5	-0.0	0.8	-1.5	0.2		0.4	0.5	0.18	-
Oman	215	219	1.0	0.0	-3.3	1.0	0.5		1.9	-1.4	-0.21	-
Qatar	153	158	5.3	0.1	-9.0	-0.4	5.5		4.0	-1.2	-0.37	-
Russian Federation	221	216	2.4	-0.0	-4.5	3.2	0.8		0.4	0.0	-0.24	-
Sweden	190	194	2.8	0.1	-0.2	-2.0	3.8		-1.6	-1.0	-0.31	-
United Arab Emirates	193	199	-2.4	-0.0	3.8	-6.1	0.7		1.6	-2.0	-0.14	-
United States	259	264	1.1	0.1	1.4	-2.9	2.6		-1.1	0.8	-0.25	D
International Avg	2447	2497	1.0	0.1	-1.2	-0.4	1.2		0.4	-0.5	-0.18	-

Keys: DIFF= Percent correct; DISC= Item discrimination; 0/A...2/C= percent in each score point for CR item or option for MC item; D= percent choosing 'D' for MC item; RDIFF= Rasch Difficulty; FLAG= Deviation from paper reported

The team at the TIMSS & PIRLS International Study Center spent two consecutive days reviewing the paperTIMSS, eTIMSS, and difference statistics for each item. Particular focus was on the difference statistics for the item review. The author prepared sets of review materials by grade and subject with three pages of statistics for each item—first paperTIMSS, followed by eTIMSS, then the difference statistics. The

team reviewed the statistics along with the paper version of each item and associated scoring guide as well as the country item documentation for any reports of deviations in formatting or translation of each digital item from its paper equivalent.

The TIMSS & PIRLS International Study Center used the results of the item review to identify any items with unusual psychometric properties (Foy et al., 2016). For example, items that are unusually easy or difficult or have negative or extremely low discrimination could be indicative of errors in implementing the item translations from the Translation System, printing, or other administration procedures. The author and another staff member at the TIMSS & PIRLS International Study Center investigated these items for possible errors by checking country item documentation and examining national instruments. Data were removed for countries and items with errors identified in both modes for subsequent analysis. This included two countries at each grade whose data were deleted due to issues with data quality caused by technical problems of assessment delivery devices and issues with the IEA data upload server. Individual item deletions were implemented for 7 countries at the fourth grade and for 5 countries at the eighth grade, with 1 to 4 items deleted for each country.

Refining the A Priori Item Hypotheses

After removing the problematic data, the TIMSS & PIRLS International Study Center used the results of the item review to re-examine the a priori item classifications and determine any items that may not be suitable for the eTIMSS environment, due to the inability of students to appropriately respond to the item as they would on paper. These items may require substantive changes for data collection that could drastically change the nature of the item possibly affecting construct equivalence. The a priori item

classifications were refined to two categories: *expected invariant* and *expected non-invariant*. *Expected invariant* items are hypothesized to be psychometrically equivalent under IRT. The TIMSS & PIRLS International Study Center deemed *expected non-invariant* items to be unsuitable for eTIMSS in their paper equivalent formats and deleted the data for further analysis.

To identify the unsuitable, *expected non-invariant* items, the author of the proposed dissertation compiled a list of all items with large differences in the percentage of missing responses between paperTIMSS and eTIMSS for review by the TIMSS & PIRLS International Study Center. A substantively larger proportion of missing responses for the eTIMSS version an item compared to the paper version was indicative that students struggled to respond to the item on PC or tablet or that the item could not be scored—either due to technical issues or the unsuitability of a digital format to assess the target construct. The list included any item with one or more country having a difference of at least 10 percent in "omitted" or "not reached" responses between paperTIMSS and eTIMSS. The team examined items flagged with five or more countries meeting these criteria for technical problems. The author examined national instruments in paper and digital formats and referenced country item documentation to check for reported differences between paper and digital versions items.

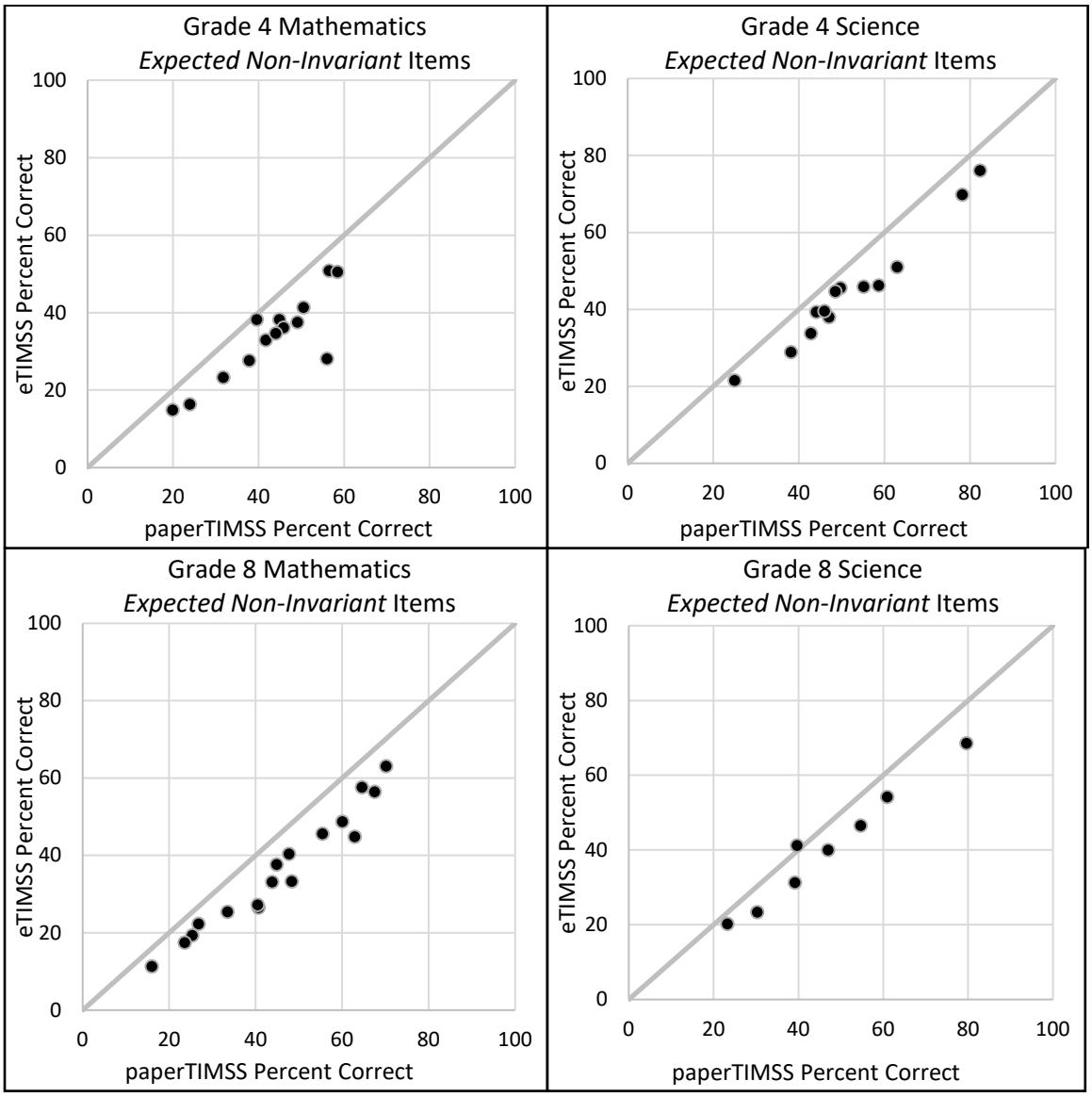
The review identified 28 items at the fourth grade (15 mathematics items and 13 science items) and 25 items at the eighth grade (17 mathematics items and 8 science items) as *expected non-invariant.* These items will require substantial changes for eTIMSS that may alter the construct measured by the item on paper. The author kept

careful documentation of issues and any required changes to the items or the eTIMSS interface to be implemented for data collection.

Among fourth grade items, the majority of *expected non-invariant* items were those that were not suitably adapted for eTIMSS and students struggled to answer. Mathematics items fitting the criteria for *expected non-invariant* mostly included items with canvas input types that could not be accurately scored and items with fraction answers that could not be input due to limitations of the eTIMSS Player at the time of the eTIMSS Pilot / Item Equivalence Study. In science, some items were adapted for eTIMSS with small type boxes—often within tables—in which students could not see their typed answer. At the eighth grade, data for multiple-choice items were deleted due to scrolling issues, which resulted in high eTIMSS omission rates. Countries' reports suggest that students may not have seen some parts of items that required substantial scrolling to find.

Exhibit 3.10 shows international average percent correct plots of the *expected non-invariant* items by grade and subject. Each point represents an item, with *x*-axis values reporting percent correct statistics based on paperTIMSS data and *y*-axis values reporting percent correct statistics based on eTIMSS data. Across both grades and subjects, all but one *expected non-invariant* item in eighth grade science was more difficult for eTIMSS than for paperTIMSS—shown by the item points falling relatively far below the identity line in Exhibit 3.10. The largest proportion of these items across the fourth and eighth grades were for mathematics.

Exhibit 3.10: Item Plots of International Average Percent Correct Statistics for *Expected Non-Invariant* Items—paperTIMSS vs. eTIMSS



Producing International Average Item Statistics

Following item review, the author computed international average item statistics by grade and subject for paperTIMSS, eTIMSS, and their differences, respectively. The author used Microsoft Excel and IBM SPSS Statistics Software Version 24 to conduct the analysis. Each country was weighted equally to compute the international average item

statistics for item difficulty, item discrimination, percent omitted, and percent not reached. The author computed standard deviations separately for each country, then pooled across countries to produce an international average standard deviation for each item statistic. Item plots produced for each grade and subject allowed for visually examining the comparability of the trend item pool based on the international average percent correct statistics. The author also produced average percent correct statistics and standard deviations by digital item type.

Item Analysis Results

The eTIMSS Item Equivalence Database at the fourth grade includes data for 21 countries and 159 items—77 mathematics items and 82 science items. At the eighth grade, 11 countries are included with data for 207 items—97 mathematics items and 110 science items. Cases were removed for students who only had valid response data in one mode of administration. Exhibit 3.11 presents the resulting sample sizes.

Exhibit 3.11: eTIMSS Item Equivalence Database—Student and Item Sample Sizes*

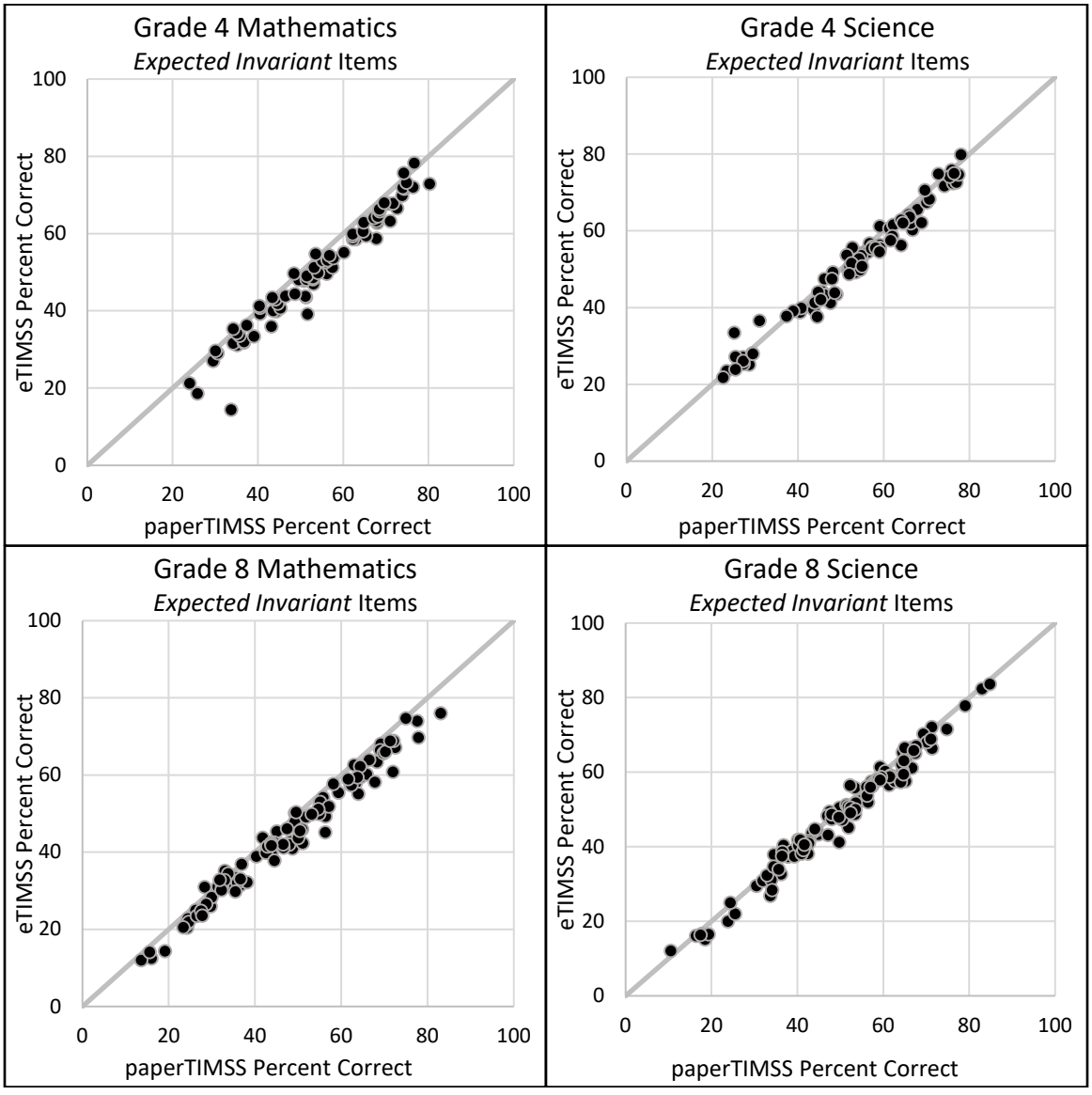
Grade	Total Cases	Mathematics		Science	
		Valid Items	Average Responses per Item	Valid Items	Average Responses per Item
Fourth Grade (21 countries)	16,894	77	4,199	82	4,200
Eighth Grade (11 countries)	9,164	97	2,278	110	2,270

* Item counts only include items classified as "expected invariant."

Exhibit 3.12 presents visual plots of the differences in item difficulty between modes, with one plot for each grade and subject. Under the within-subjects, counterbalanced research design for the eTIMSS Pilot / Item Equivalence Study, items

with similar measurement properties across modes should have very similar percent correct values and show only small, random deviations from the identity line. Overall, points fall close to the identity line, and appear to have small, random deviations from the identity line throughout the range of percent correct values, suggesting that items are generally equally difficult between modes, on average.

Exhibit 3.12: Item Plots of International Average Percent Correct Statistics for *Expected Invariant* Items—paperTIMSS vs. eTIMSS



However, the results provide evidence of a general mode effect for the TIMSS trend items. Overall, on average, items were more difficult for eTIMSS than for paperTIMSS. Particularly for mathematics, the plots in Exhibit 3.12 show most points clustering below the identity line, indicating the items generally were more difficult (i.e., had smaller percent correct statistics) for eTIMSS than for paperTIMSS. The plots for mathematics at the fourth and eighth grades are similar, with more points falling below the identity line than above.

The plots for science show more equal distributions of points around the identity line, suggesting the mode effect may be smaller than for mathematics. Interestingly, the borderline outlier items in fourth grade science fall above the identity line and were easier for students on eTIMSS compared to paper. The plot for eighth grade science items shows the most similarity between paperTIMSS and eTIMSS item statistics with an approximately equal distribution of points above and below the identity line.

Comparing the plots in Exhibit 3.12 with those of *expected non-invariant* items in Exhibit 3.10 provides confirmation of the refined item classifications. Not only were the *expected non-invariant* items not well adapted to eTIMSS and require more substantial changes from their paper formats, but these items also show much larger mode effects on item difficulty compared to the *expected invariant* items shown below.

Exhibit 3.13 presents the international average percent correct statistics for each grade and subject. Fourth grade mathematics items showed the largest average difference in item difficulty between paperTIMSS and eTIMSS, with a 3.6 percent difference in average percent correct statistics between modes. This indicates that internationally, on average, 3.6 percent more fourth grade students answered paperTIMSS mathematics

items correctly compared to eTIMSS mathematics items. Eighth grade mathematics items showed a similar effect, with a 3.4 percent average difference in percent correct between paperTIMSS and eTIMSS.

Science items at both grades showed smaller effects of the mode of administration on item difficulty compared to mathematics items. On average across the fourth grade science items, 1.7 percent more students answered the items correctly on paper compared to eTIMSS. Eighth grade science items showed to be the most similar in terms of item difficulty, with an average percent correct difference of only 1.5 percent.

Exhibit 3.13: International Average Percent Correct Statistics (Item Difficulty)

Grade/Subject	Valid N	International Average Percent Correct (Item Difficulty)		
		paperTIMSS	eTIMSS	Difference
Fourth Grade (21 countries)				
Mathematics	77	53.7 (18.1)	50.1 (18.0)	3.6 (6.0)
Science	82	53.1 (17.9)	51.3 (17.5)	1.7 (6.0)
Eighth Grade (11 countries)				
Mathematics	97	47.4 (19.1)	44.0 (18.5)	3.4 (5.3)
Science	110	49.6 (18.4)	48.1 (18.5)	1.5 (5.7)

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

Exhibit 3.14 presents international average item discrimination statistics across the trend items by grade and subject. The results suggest there was little to no effect of mode of administration on item discrimination statistics, with less than 0.03 average difference in point-biserial correlation coefficients for each subject and grade, and with little variation across items and countries ($SD < 0.10$).

Exhibit 3.14: International Average Point-Biserial Correlations (Item Discrimination)

Grade/Subject	Valid N	International Average Point-Biserial Correlation (Item Discrimination)		
		paperTIMSS	eTIMSS	Difference
Fourth Grade (21 countries)				
Mathematics	77	0.42 (0.12)	0.41 (0.12)	0.01 (0.09)
Science	82	0.37 (0.10)	0.36 (0.11)	0.00 (0.09)
Eighth Grade (11 countries)				
Mathematics	97	0.42 (0.13)	0.41 (0.14)	0.02 (0.03)
Science	110	0.37 (0.12)	0.37 (0.12)	0.01 (0.09)

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

Exhibit 3.15 reports international average percent omitted statistics by each grade and subject. The negative percentages in the difference column for fourth grade mathematics items and for the two pools of eighth grade items indicate that a larger percentage of students did not answer eTIMSS items compared to paperTIMSS items, on average. These differences were small—less than 1 percent—suggesting that, overall, students did not struggle to respond to the digital versions of the *expected invariant* items more than the paper versions.

Exhibit 3.15: International Average Percent Omitted Statistics

Grade/Subject	Valid N	International Average Percent Omitted		
		paperTIMSS	eTIMSS	Difference
Fourth Grade (21 countries)				
Mathematics	77	4.1 (4.1)	4.8 (5.1)	-0.7 (4.2)
Science	82	4.9 (5.7)	4.7 (4.8)	0.2 (3.6)
Eighth Grade (11 countries)				
Mathematics	97	4.1 (5.1)	4.4 (5.4)	-0.3 (2.7)
Science	110	4.5 (6.2)	4.6 (6.7)	-0.1 (3.4)

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

The percent not reached statistics in Exhibit 3.16 were of interest to make sure students did not take longer to answer the trend items on a digital device than on paper. The results suggest there was no mode effect on these percentages across items.

Exhibit 3.16: International Average Percent Not Reached Statistics

Grade/Subject	Valid N	International Average Percent Not Reached		
		paperTIMSS	eTIMSS	Difference
Fourth Grade (21 countries)				
Mathematics	77	1.0 (1.1)	1.0 (1.4)	0.0 (1.2)
Science	82	1.1 (1.4)	1.6 (2.0)	-0.4 (1.6)
Eighth Grade (11 countries)				
Mathematics	97	0.8 (1.2)	0.7 (1.2)	0.0 (0.9)
Science	110	0.4 (0.5)	0.5 (0.8)	-0.1 (0.6)

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

Results by Digital Item Type

This section presents the international average percent correct statistics by digital item type for each grade and subject. As discussed in Chapter 2, these results were informative for developing items for the Field Test. The results should be interpreted with caution due to the small sample sizes.

Exhibit 3.17 presents the international average percent correct statistics by digital item type for fourth grade mathematics items. In both paper and digital modes, students performed best on multiple-choice items, with 58.3 percent correct on paperTIMSS and 55.5 percent correct on eTIMSS, on average. As anticipated, the six remaining canvas items were the most difficult item type for fourth grade students, with 35.0 percent correct for eTIMSS.

Items requiring students to use the keyboard showed the largest advantage for paper items, with 7.8 percent more students answering these items correctly on paper

compared to eTIMSS, on average. Across items with number pad and canvas inputs, respectively, there was a 4.4 percent correct advantage for paper items. Multiple-choice and compound multiple-choice items showed the smallest mode effects, with each type showing average differences around 3 percent.

Exhibit 3.17: International Average Percent Correct Statistics (Item Difficulty) by Digital Item Type—Fourth Grade, Mathematics

Digital Item Type	Valid N	International Average Percent Correct (Item Difficulty)		
		paperTIMSS	eTIMSS	Difference
Multiple-Choice	42	58.3 (17.4)	55.5 (17.0)	2.8 (5.5)
Compound Multiple-Choice	2	44.3 (16.2)	41.1 (14.7)	3.2 (4.5)
Keyboard	2	51.2 (13.9)	43.5 (14.4)	7.8 (9.4)
Number Pad	25	50.5 (16.3)	46.1 (16.7)	4.4 (6.3)
Canvas	6	39.4 (21.3)	35.0 (19.6)	4.4 (6.4)
Total	77	53.7 (18.1)	50.1 (38.1)	3.6 (6.0)

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

The results in Exhibit 3.18 show average percent correct statistics by digital item type for fourth grade science items. Similar to the results for mathematics, digital multiple-choice items in science were the least difficult compared to digital compound multiple-choice and keyboard item types (56.7% vs. 47.6% and 45.2%, respectively).

The fourth grade science items were found to behave more similarly across the variations in digital item types compared to mathematics. Multiple-choice items showed an average difference of 2.0 percent, compound-multiple choice items showed an average difference of 1.9 percent, and keyboard items showed an average difference of 1.4 percent. This result was positive and suggests that students' typed responses were not scored more harshly than written responses, as the research suggests could occur (Horkay et al., 2006; Russell, 2002).

Exhibit 3.18: International Average Percent Correct Statistics (Item Difficulty) by Digital Item Type—Fourth Grade, Science

Digital Item Type	Valid N	International Average Percent Correct (Item Difficulty)		
		paperTIMSS	eTIMSS	Difference
Multiple-Choice	43	58.7 (14.6)	56.7 (14.4)	2.0 (5.2)
Compound Multiple-Choice	3	49.5 (27.6)	47.6 (27.1)	1.9 (5.4)
Keyboard	36	46.6 (18.7)	45.2 (18.3)	1.4 (6.7)
Number Pad	0	- -	- -	- -
Canvas	0	- -	- -	- -
Total	82	53.1 (17.9)	51.3 (17.5)	1.7 (6.0)

( ) Standard deviations appear in parentheses. Because of rounding some results may appear inconsistent.
A dash (-) indicates comparable data not available.

Exhibit 3.19 presents the international average statistics for percent correct by input type for the eighth grade mathematics items. As expected in the a priori analysis, number pad items showed the largest average mode effect for item difficulty, with 3.7 percent more students answering these items correct on paperTIMSS than on eTIMSS. Multiple-choice items and keyboard items had similar differences, with each having 3.2 percent more students answering these items correctly on paper compared to eTIMSS. However, students performed better on digital multiple-choice items compared to keyboard items. Keyboard items showed an international average percent correct of 26.0 percent correct for eTIMSS, whereas 49.2 percent of students answered multiple-choice items correctly on eTIMSS.

Exhibit 3.19: International Average Percent Correct Statistics (Item Difficulty) by Digital Item Type—Eighth Grade, Mathematics

Digital Item Type	Valid N	International Average Percent Correct (Item Difficulty)		
		paperTIMSS	eTIMSS	Difference
Multiple-Choice	61	52.4 (18.1)	49.2 (17.1)	3.2 (5.4)
Compound Multiple-Choice	0	- -	- -	- -
Keyboard	7	29.2 (17.8)	26.0 (17.4)	3.2 (4.4)
Number Pad	29	41.1 (16.9)	37.4 (16.1)	3.7 (5.2)
Canvas	0	- -	- -	- -
Total	97	47.4 (19.1)	44.0 (18.3)	3.4 (5.3)

( ) Standard deviations appear in parentheses. Because of rounding some results may appear inconsistent.
A dash (-) indicates comparable data not available.

Exhibit 3.20 presents the international average item difficulty statistics by digital item type for eighth grade science items. Similar to the results at the fourth grade, eighth grade science items showed less variation in student performance across the digital item types compared to mathematics. Multiple-choice items were the least difficult among eighth grade science items, on average. Over half the students answered multiple-choice items correctly in both paperTIMSS and eTIMSS (55.4% and 54.1%, respectively). Keyboard items showed to be the most difficult input type for eighth grade science, with 39.6 percent of students answering these correctly for eTIMSS, on average.

The two number pad items assessing eighth grade science showed the largest differences in percent correct statistics between modes, with 3.3 percent more students answering paperTIMSS items correctly compared to eTIMSS items, on average. Keyboard items showed an international average difference in item difficulty of 1.7 percent favoring paperTIMSS. Multiple-choice and compound multiple-choice items exhibited the smallest effects on item difficulty, with international average differences of 1.3 percent and 1.1 percent, respectively.

Exhibit 3.20: International Average Percent Correct Statistics (Item Difficulty) by Digital Item Type—Eighth Grade, Science

Digital Item Type	Valid N	International Average Percent Correct (Item Difficulty)		
		paperTIMSS	eTIMSS	Difference
Multiple-Choice	57	55.4 (16.3)	54.1 (16.2)	1.3 (5.1)
Compound Multiple-Choice	7	52.9 (23.3)	51.8 (23.2)	1.1 (4.4)
Keyboard	44	41.4 (17.6)	39.6 (17.7)	1.7 (6.5)
Number Pad	2	52.0 (10.0)	48.7 (10.2)	3.3 (4.0)
Canvas	0	- -	- -	- -
Total	110	49.6 (18.4)	48.1 (18.5)	1.5 (5.7)

( ) Standard deviations appear in parentheses. Because of rounding some results may appear inconsistent.
A dash (-) indicates comparable data not available.

Item Analysis Summary

Although there were trend items with no difference in percent correct between modes, the results of the item analysis suggest a definite mode effect across the TIMSS trend items. On average across countries, paper versions of items were less difficult than their digital counterparts. The mode effects were larger for mathematics items compared to science items. Fortunately, negligible differences in item discrimination statistics between modes at each grade suggest no effect of eTIMSS on the TIMSS mathematics and science constructs measured by the paper instruments—only item difficulties showed differences (Winter, 2010). There were also little or no effects on the percentage of "not reached" responses across items, indicating that the testing times allotted for the paperTIMSS assessment were adequate for eTIMSS.

Differential rates of item omissions by mode were specific to items not well adapted to the eTIMSS interface at the time of Pilot / Item Equivalence Study, and may undergo further adaptation for data collection in 2019. Among the remaining constructed response items, students tended to omit digital items more often than paper items, but

these differences were small. In some instances, students omitted multiple-choice items less often in their digital formats compared to paper.

The results by digital item type indicated that there were differential mode effects on percent correct statistics depending on the response action required to answer each item. This is consistent with previous research that differences in responses requirements to items could influence mode effects (Duque, 2016). It is important to note that small sample sizes and large variation in the values limit the interpretability of the item analysis results by digital item type. Nevertheless, the results were informative for developing new items for TIMSS 2019. In mathematics as well as in eighth grade science, canvas and number pad items were the most different across modes in terms of difficulty, with paperTIMSS formats being less difficult than the eTIMSS formats. For the Field Test, there were no canvas item types and the number pad was improved. Fourth grade science items showed little to no differentiation by item type in terms of mode differences.

Due to the evidence for an overall mode effect, the TIMSS & PIRLS International Study Center and Educational Testing Service decided that further analysis of item equivalence was not warranted. Had there been no differences in classical item statistics between paperTIMSS and eTIMSS, an IRT approach to examining item equivalence would allow for isolating any effect of mode of administration on the IRT item parameters (for examples of this approach see Buerger, Kroehne, & Goldhammer, 2016; Oliveri & von Davier, 2011; 2014).

Phase 3: Scale Score Analysis

Methodological Overview

Given that the item-level analysis showed a general mode effect, Phase 3 of the eTIMSS Pilot / Item Equivalence Study investigated the effect of the computer- and tablet-based administration on TIMSS mathematics and science achievement scores at each grade. Through the procedures described in the subsections below, ETS simulated the results of the mode effect by replicating the TIMSS scaling methodology (see Foy & Lin, 2016; Martin, Mullis, Foy, & Hooper, 2016a) on the eTIMSS Pilot / Item Equivalence Study data. Item parameters were first estimated for paperTIMSS, and the resulting paperTIMSS parameters were used to estimate achievement scores for both the paperTIMSS data and eTIMSS data. By fixing the item parameters to those based on the paperTIMSS results, the mode effect was captured by the differences in group means between paperTIMSS and eTIMSS. This approach provided a way of understanding the mode effects that would occur on the proficiency distributions without adjusting for them in the analysis. Commonly accepted criteria of score comparability include: 1) score distributions being approximately the same; and 2) individuals—or subgroups in the TIMSS case—being rank ordered in approximately the same way (APA, 1986; DePascale et al., 2016; Winter, 2010).

The matrix sampling design described earlier in this chapter means that students received only a subset of the entire assessment pool, so individual student scores do not reliably measure mathematics and science achievement as defined and measured by TIMSS. To produce more accurate score estimates for populations and subpopulations of students, TIMSS uses plausible values methodology with conditioning (Martin, Mullis,

Foy, & Hooper, 2016a; Mislevy, 1991). Using this approach, TIMSS estimates five imputed proficiency scores called "plausible values" for each student based on their estimated ability distribution and conditioned upon student and class characteristics. Conducting analysis across all five plausible values allows for more accurate estimation of population and subpopulation parameters and the level of uncertainty around the estimates.

Assessing the comparability of the resulting score estimates for paperTIMSS and eTIMSS involved examining the two score distributions by grade and subject. Mean scores by country were examined to determine whether country rankings differed across modes. The within-subjects design of the eTIMSS Pilot / Item Equivalence Study allowed for the use and interpretation of correlational analyses of scale scores to provide evidence of score and construct comparability (Buerger et al., 2016; Winter, 2010). Cross-mode correlation coefficients between paperTIMSS and eTIMSS score distributions were produced, corrected for the reliability of the country means, and examined for construct equivalence. The author of this dissertation used the plausible values produced by ETS to replicate the results and extend their analysis.

Scale Score Analysis Procedure

Calibrating the Items

Similar to TIMSS scaling methodology, item parameters and student achievement estimates were estimated by staff at ETS using mixed (2- and 3-parameter) IRT models based on the international sample of responses, with each country's response data contributing equally to calibration (Foy & Lin, 2016; Martin, Mullis, Foy, & Hooper, 2016a). ETS staff conducted item calibration separately by grade and subject using

PARSCALE software (Muraki & Bock, 1991). Senate weights (SENWGT) were used to weight cases, which provides each country an equal weight of 500 for calibration. First, paperTIMSS item parameters were estimated based on responses to paper items, with multiple-choice item parameters estimated using a three parameter (3-PL) IRT model and 1-point constructed response items estimated using a two parameter (2-PL) IRT model. Two- and three- parameter IRT models give the probability that a student with proficiency *k,* characterized by the unobserved θ_k, will respond correctly to item i as follows:

$$P(x_i = 1 \mid \theta_k, a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + \exp(-1.7 \cdot a_i \cdot (\theta_k - b_i))}, \tag{3.1}$$

where x_i is the response to item i (1 = correct, 0 = incorrect); θ_k is the student's unobserved proficiency on scale k; a_i is the discrimination of item i; b_i is the difficulty, or location parameter of item i; and c_i is the guessing parameter, or lower asymptote parameter of item i. In the two-parameter model, c_i is fixed at zero.

Item parameters for paperTIMSS constructed response items worth up to 2 score points were estimated using a generalized partial credit model (Muraki, 1992). This models the probability that a student with θ_k on scale k will score in the l^{th} category on item i as follows:

$$P(x_i = l \mid \theta_k, a_i, b_i, d_{i,1}, ..., d_{i,m_i-1}) = \frac{\exp\left(\sum_{v=0}^{t} 1.7 \cdot a_i \cdot (\theta_k - b_i + d_{i,v})\right)}{\sum_{g=0}^{m_i-1} \exp\left(\sum_{v=0}^{g} 1.7 \cdot a_i \cdot (\theta_k - b_i + d_{i,v})\right)}, \tag{3.2}$$

where m_i is the number of response categories for item i; x_i is the response to item i (0, 1, or 2); and $d_{i,l}$ is the threshold parameter for category l.

Estimating Scale Scores

Next, ETS used DGROUP software (Rogers, Tang, Lin, & Kandathil, 2006) to estimate student proficiency scores separately by grade for paperTIMSS and eTIMSS, respectively, using the item parameters estimated from the paperTIMSS calibration. With a bi-variate conditioning model, mathematics and science proficiency scores were estimated concurrently for each student based on both mathematics and science item response data as well as background variables for conditioning. Five plausible values were drawn for each student *j* in both mathematics and science from the conditional distribution as follows:

$$P(\theta_j \mid x_j, y_j, \Gamma, \Sigma) \propto P(x_j \mid \theta_j) \cdot P(\theta_j \mid y_{j,} \Gamma, \Sigma), \tag{3.3}$$

where θ_j is a vector of mathematics and science scale values; $P(x_j \mid \theta_j)$ is the product over the mathematics and science scales of the independent likelihoods given by responses to items within each scale, based on the fixed paperTIMSS item parameters; and $P(\theta_j \mid y_{j,} \Gamma, \Sigma)$ is the bi-variate joint density of mathematics and science proficiencies. The proficiencies are conditional upon students' observed values y_j on the background variables and parameters Γ (regression coefficients or principle components of the background variables for each country) and Σ (common variance).[1]

Conditioning variables included variables from the student questionnaire (gender, number of books in the home, access to a computer or tablet at school, and three variables about computer experience, as well as parents' education for eighth grade students), class

[1] For more detail on the model and procedures for estimating student proficiency in TIMSS, see Martin, Mullis, Foy, and Hooper (2016a). See Mislevy (1991) for theoretical background.

mean expected a posterior (EAP) scores, country, and the interactions between country and each other conditioning variable.

ETS transformed the resulting plausible values from the resulting theta metric to an approximate TIMSS scale, which has a mean of 500 and a standard deviation of 100. First, ETS used the TIMSS 2015 IRT item parameters (Foy & Lin, 2016) to apply a Stocking-Lord transformation (Stocking & Lord, 1983) to the resulting item characteristic curves, placing them on the TIMSS 2015 theta scale. Then, the same linear transformation constants that were used to transform the TIMSS 2015 scores onto the TIMSS reporting metric were applied (see Foy & Lin, 2016), resulting in student proficiency scores on the TIMSS reporting scale (500,100).

Producing Statistics to Examine Score Comparability

The author of this dissertation replicated the international average results produced by ETS and extended their analysis. Analysis was conducted separately for the fourth grade and eighth grade for mathematics and science, respectively. To produce the results in the next section, the author used IBM SPSS Statistics Software Version 24 and scripts for SPSS created by IEA Hamburg for use with the IEA's IDB Analyzer software. First, a SENWGT was computed that weighted each country equally in the analysis. For each country, the SENWGT was produced by dividing 500 by the total number of students so that the weights sum to 500 for each country.

For each grade and subject, international average scale scores, standard deviations, and standard errors were computed for paperTIMSS and eTIMSS, respectively. The IEA's IDB Analyzer software script for SPSS applied the SENWGT, computed the average of each plausible value across all cases in the database, and

aggregated the results across the plausible values for interpretation. The script also produced jackknife standard errors for each country that account for both measurement and sampling error with the jackknife repeated replication method (Foy & LaRoche, 2016; Rust, 2014). The stratification variables used to produce the jackknife standard errors were based only on school membership within each country, so the standard errors account for the clustering of the students in schools. Because the student samples were not drawn randomly, the standard errors are not an accurate reflection of the population data. However, the standard errors are more errors are more realistic than without the jackknifing applied and helped in interpreting the results.

In addition to paperTIMSS scores and eTIMSS scores, the author computed international average difference scores by subtracting each eTIMSS plausible value from its respective paperTIMSS plausible value for each case, then following the same steps to produce the average difference scores, standard deviations, and standard errors.

To aid in interpreting the magnitude of the score differences, effect sizes were calculated using Cohen's d_{rm} for repeated measures (Cohen, 1988), which accounts for the size of the correlation between paperTIMSS and eTIMSS scores:

$$d_{rm} = \frac{M_{diff}}{\sqrt{SD_{paper}^2 + SD_{eTIMSS}^2 - 2 \cdot r \cdot SD_{paper} \cdot SD_{eTIMSS}}} \cdot \sqrt{2 \cdot (1 - r)} \qquad (3.4)$$

To examine the distribution and relative magnitudes of the difference scores across countries, the author produced bar charts for the mean difference between paperTIMSS and eTIMSS scores by country. To better interpret the magnitude of the score differences across countries, the charts include 95 percent confidence intervals for each mean score difference based on the approximate jackknife standard errors.

Lastly, ETS computed cross-mode correlations to assess the equivalence of the mathematics and science constructs for paperTIMSS and eTIMSS. Interpretation of cross-mode correlation coefficients may be limited without adjusting for the reliability of the scale scores (Winter, 2010). Therefore, ETS used approximate standard errors in computing correlation coefficients between paperTIMSS and eTIMSS scores.

Scale Score Analysis Results

The research design and the methodology implemented to estimate the proficiency scores isolated the effect of mode of administration on the TIMSS achievement scores. If there were no mode effects, the results would look similar across paperTIMSS and eTIMSS scores within each country and overall. Note that the results are not generalizable to the full TIMSS population or population subgroups.

Exhibit 3.21 presents the international average scale scores for each mode by grade and subject. Overall, scores based on paperTIMSS items are higher than scores based on eTIMSS items, providing evidence that the TIMSS assessments are more difficult using the digital administration. The effect was larger for mathematics than for science. At the fourth and eighth grades, the results show an international average difference of 14 points favoring paper for mathematics scores. Science showed a smaller effect with an average difference of 8 score points at the fourth grade and 7 score points at the eighth grade. Variance in mean scores were approximately equal across modes, on average. However, the large standard deviations reported for the difference values suggest that the magnitude of mode effects differed substantially across students.

Exhibit 3.21: International Average Scale Scores and Standard Deviations

Grade/Subject	International Average Achievement Scores		
	paperTIMSS	eTIMSS	Difference
Fourth Grade (21 countries)			
Mathematics	527 (86)	513 (86)	14 (40)
Science	526 (86)	518 (86)	8 (41)
Eighth Grade (11 countries)			
Mathematics	511 (98)	497 (97)	14 (42)
Science	521 (92)	514 (90)	7 (43)

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

To aid in interpreting the magnitude of the international average difference sores, Exhibit 3.22 presents the same results from Exhibit 3.21 but with approximate international average standard errors for each score. The standard errors are about the same for paperTIMSS and eTIMSS, providing support for score comparability. The larger magnitudes at the eighth grade may be attributed to the smaller sample sizes.

Exhibit 3.22: International Average Scale Scores, Standard Errors, and Mode Effect Sizes

Grade/Subject	International Average Achievement Scores			Mode Effect Size (d_{rm})
	paperTIMSS	eTIMSS	Difference	
Fourth Grade (21 countries)				
Mathematics	527 (1.4)	513 (1.4)	14 (0.7)	0.16
Science	526 (1.5)	518 (1.5)	8 (0.6)	0.09
Eighth Grade (11 countries)				
Mathematics	511 (2.2)	497 (2.3)	14 (1.0)	0.14
Science	521 (2.6)	514 (2.5)	7 (1.0)	0.08

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

Exhibit 3.22 also includes the effect size for each score difference, computed using the formula in equation (3.4). The mode effects are all considered small ($d_{rm} < 0.02$) based on Cohen's benchmarks for interpreting effect sizes (Cohen, 1988)—where

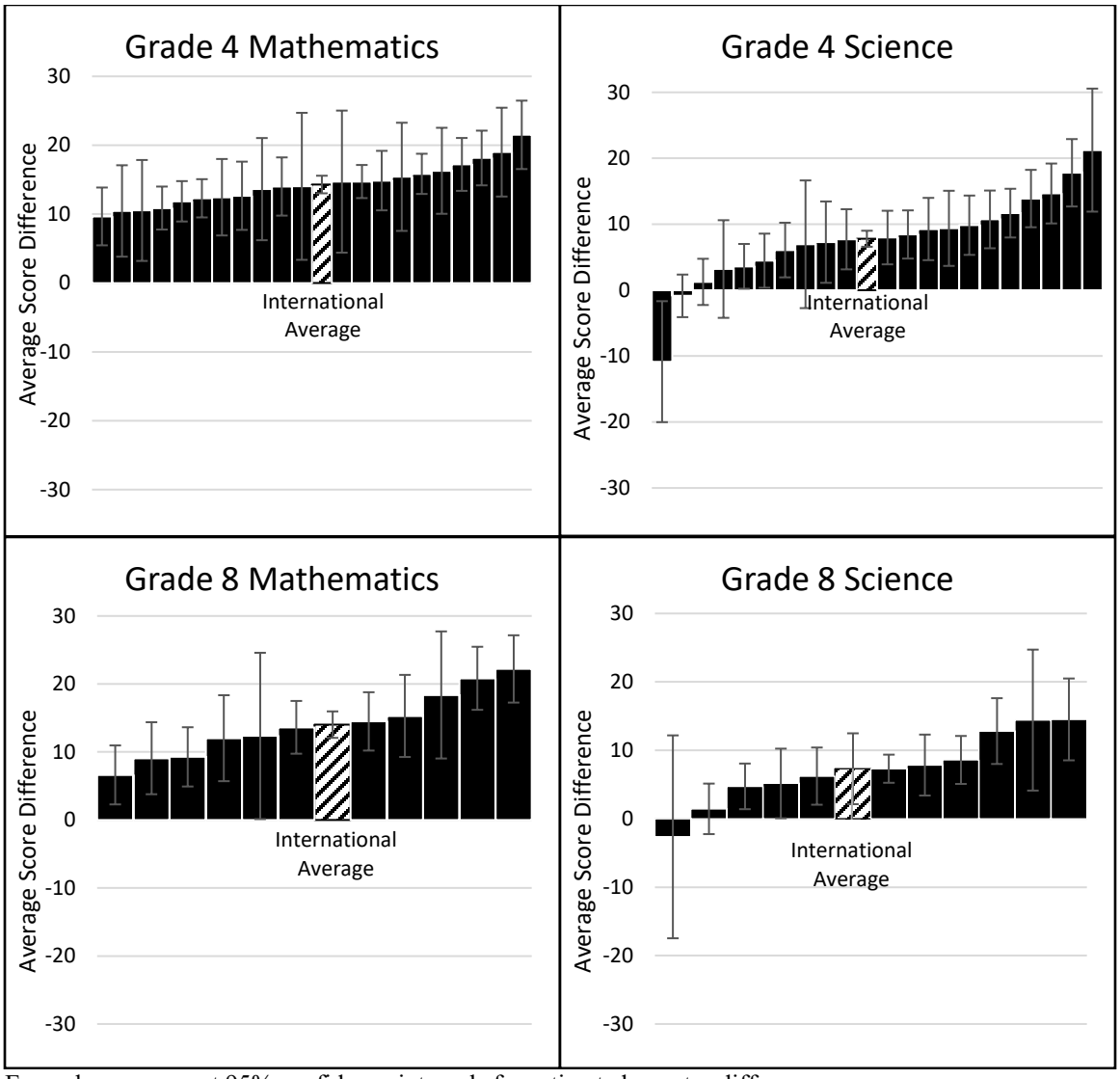
d_{rm} = 0.16 for fourth grade mathematics, 0.09 for fourth grade science, 0.14 for eighth grade mathematics, and 0.08 for eighth grade science. However, this effect size measure is conservative, especially with large sample sizes and highly correlated scores (Lakens, 2013).

Overall, the effect sizes are similar in magnitude to those found in earlier mode effect studies (DePascale et al., 2016; Wang, Jio, Young, Brooks, & Olson, 2007; Way et al., 2016). Despite some variation in mode effect sizes in the literature, most are small in magnitude (Cohen, 1988). Earlier studies of mode effects in NAEP mathematics found effects of 0.015 and 0.14 standard deviations between computer-based and paper-based test performance (Poggio, Glasnapp, Yang, & Poggio, 2005; Bennett et al., 2008). Larger meta-analyses have computed average effect sizes of 0.06 to 0.07 in mathematics (Kingston, 2008; Wang et al., 2007) and 0.08 in science (Kingston, 2008). Also, a recent study of three countries' data from the PISA 2015 field trial found average score differences of 14 score points in mathematics and 15 score points in science between paper and digital modes (Jerrim, 2018), similar to the mode effects found for TIMSS mathematics.

Exhibit 3.23 presents bar charts for each subject by grade illustrating the distribution of average scale score differences across countries. The bars are ordered by the magnitude of the average score difference. Error bars reflect 95 percent confidence intervals of the difference scores. The graphs for mathematics at both grades show that all countries performed better on paperTIMSS compared to eTIMSS in mathematics, on average. The graphs for science show more variation across countries in the size of the mode effect, with most countries having higher mean scores for paperTIMSS compared

to eTIMSS, but with some countries performing better on eTIMSS or similar across modes, on average.

Exhibit 3.23: Country Distribution of Average Scale Score Differences between paperTIMSS and eTIMSS



Errors bars represent 95% confidence intervals for estimated country difference scores.

However, much of the variance in country difference scores may be negligible after accounting for the uncertainty of the estimates with regard to sampling error. The size of 95 percent confidence intervals suggest that most of the score differences across

countries are relatively equal within error. The exceptions may be the countries exhibiting negative differences in science scores, indicating better eTIMSS performance compared to paperTIMSS. However, at the eighth grade, the negative science difference score also is the least precise, so could be positive in magnitude with a nationally representative sample.

Exhibit 3.24 presents the international average cross-mode correlation coefficients (adjusted for reliability) for the relationships between paperTIMSS and eTIMSS scores by grade and subject. The results show strong relationships between the constructs across modes ($r > 0.95$), providing evidence for construct equivalence between paper and digital modes. This suggests that the mode of administration did not have an effect on the TIMSS mathematics and science constructs. The relationships were approximately equal by grade and subject, suggesting that the overall effect of the digital mode on the TIMSS score distributions are similar.

Exhibit 3.24: International Average Cross-Mode Correlation Coefficients (Adjusted for Reliability)

Grade	International Average Cross-Mode Correlation Coefficient (*r*)	
	Mathematics	**Science**
Fourth Grade (21 countries)	0.96	0.96
Eighth Grade (11 countries)	0.97	0.96

Scale Score Analysis Summary

The results of the scale score analysis confirm an overall mode effect on TIMSS achievement scores, with stronger mode effects exhibited by mathematics scores compared to science scores. The overall differences in international average scores

between paperTIMSS and eTIMSS did not vary across grades for each subject. Score distributions were approximately the same across modes, but with more variation in country mean difference scores for science compared to mathematics. However, the difference scores are mostly within the approximated margin of error.

Despite the presence of mode effects, the results suggest that the mathematics and science constructs measured by the trend items are the same in paperTIMSS and eTIMSS and that the two scores are comparable. Examining mean scores by country (not reported in this dissertation) indicated that the ordering of country mean scores did not differ between paperTIMSS and eTIMSS at the high and low ends of the score distributions, and differed only somewhat toward the middle. Additionally, the magnitudes of the cross-mode relationships after correcting for reliability suggest strong construct equivalence between modes. There were approximately equal cross-mode relationships across the grades and subjects, suggesting that the magnitude of differences in item difficulties and scale scores between modes did not strongly affect the resulting score distributions.

Discussion of the Results of the eTIMSS Pilot / Item Equivalence Study

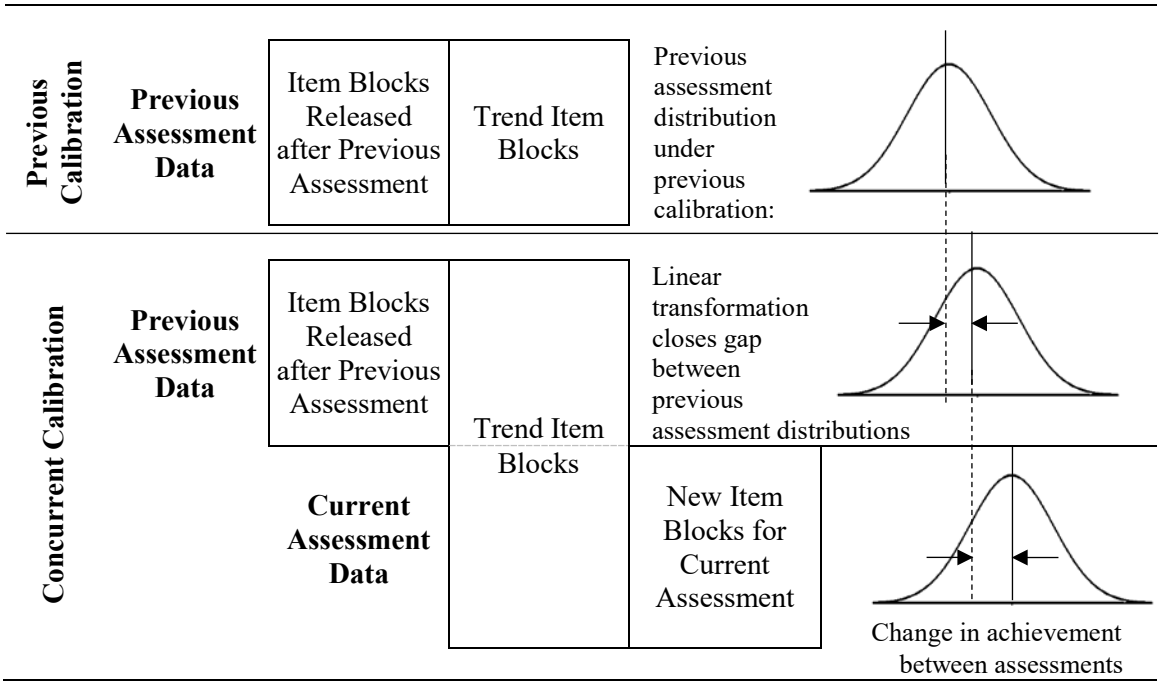
Taken together, the results of the eTIMSS Pilot / Item Equivalence Study provide a foundation for the steps necessary to ensure TIMSS scale scores are comparable between modes following data collection in 2019. The first qualitative phase of item analysis provided evidence that the majority of the trend items are face equivalent in paper and eTIMSS formats. Although the item analyses in Phase 2 showed that the

difficulties of the trend items were affected by the new eTIMSS administration, the results suggest that, overall, the TIMSS mathematics and science constructs were unaffected by the transition to eTIMSS. Therefore, the difference in scores that resulted from the mode effects on item difficulties can be corrected through equating (Winter, 2010).

Measuring Trends in TIMSS

The typical concurrent calibration process TIMSS follows to link scales between subsequent assessments involves adjusting for small differences in trend item parameters between cycles through a linear transformation that aligns the distribution of scores with the distribution of the previous cycle. This process is illustrated in Exhibit 3.25.

Exhibit 3.25: Concurrent Calibration Model Used for TIMSS Trend Measurements



Source: Adapted from Foy & Lin (2016).

For example, TIMSS 2015 item parameters were estimated based on the combined achievement data from both TIMSS 2011 and TIMSS 2015 assessments (Foy

& Lin, 2016). A linear transformation removed the gap in the distribution of the TIMSS 2011 data under the TIMSS 2015 concurrent calibration to match the distribution of the same data based on the TIMSS 2011 calibration. The linear transformations were applied to the plausible values with:

$$PV^*_{k,i} = A_{k,i} + B_{k,i} \cdot PV_{k,i}, \tag{3.5}$$

where $PV_{k,i}$ is the TIMSS 2015 plausible value *i* of scale *k* before the transformation; $PV^*_{k,i}$ is the TIMSS 2015 plausible value *i* of scale *k* after transformation; and $A_{k,i}$ and $B_{k,i}$ are linear transformation constants. The linear transformation constants were computed with the formulas in equations (3.6) and (3.7) based on plausible values computed from both TIMSS 2011 and TIMSS 2015 scores:

$$B_{k,i} = \frac{\sigma_{k,i}}{\sigma^*_{k,i}} \tag{3.6}$$

$$A_{k,i} = \mu_{k,i} + B_{k,i} \cdot \mu^*_{k,i}, \tag{3.7}$$

where $\mu_{k,i}$ and $\sigma_{k,i}$ are the international mean and standard deviation of scale *k* based on plausible value *i* from TIMSS 2011 and $\mu^*_{k,i}$ and $\sigma^*_{k,i}$ are the international mean and standard deviation of scale *k* based on plausible value *i* from the TIMSS 2011 data based on the TIMSS 2015 concurrent calibration. After applying the linear transformation to the TIMSS 2015 scores, the score differences that remain between assessments reflect the change in student achievement over time.

Trend item parameters typically change only slightly between assessments, usually due to the presence of new assessment items and some differences in the pool of trend countries between cycles. The large pool of common items (trend items) and

common countries keep these fluctuations relatively small. However, the results of the eTIMSS Pilot / Item Equivalence Study showed that the item parameters between paperTIMSS and eTIMSS will not be similar enough for the usual concurrent calibration approach—particularly with a 14 point average difference for mathematics.

In the TIMSS context, a difference of 14 points is substantial and corresponds to one third of the approximate 60-point difference constituting a grade level in the primary grades and half of the approximate 30-point difference constituting a grade level in middle school (Martin, Mullis, Beaton, Gonzalez, Smith, & Kelly, 1998; Mullis, Martin, Beaton, Gonzalez, Kelly, & Smith, 1998). These international average differences in mathematics and science achievement between consecutive grade-levels from TIMSS 1995 are similar to more recent results from TIMSS 2015 (Martin, Mullis, Foy, & Hooper, 2016b; Mullis, Martin, Foy, & Hooper, 2016). Norway participated in TIMSS 2015 with two grade-levels of students taking the fourth and eighth grade assessments, respectively. Between grades 4 and 5, Norway had a 56-point difference in mathematics (493 vs. 549) and a 45-point difference in science (493 vs. 538). Between grades 8 and 9, Norway had a 25-point difference in mathematics (487 vs. 512) and a 20-point difference in science (489 vs. 509).

A difference of 14 score points also is substantial in the context of trend results between subsequent TIMSS assessments. TIMSS is designed to have a standard error no greater than 3.5 percent of the standard deviation associated with each country's mean achievement score (LaRoche, Joncas, & Foy, 2016). The TIMSS reporting scale has a standard deviation of 100, so student samples should provide for a standard error of 3.5 points. This corresponds to a 95 percent confidence interval of ±7 score points for an

achievement mean and ±10 score points for the difference between means from subsequent assessment cycles. Therefore, a 14-point difference would constitute a significant difference between mean scores and must be corrected to ensure accurate trend measurements.

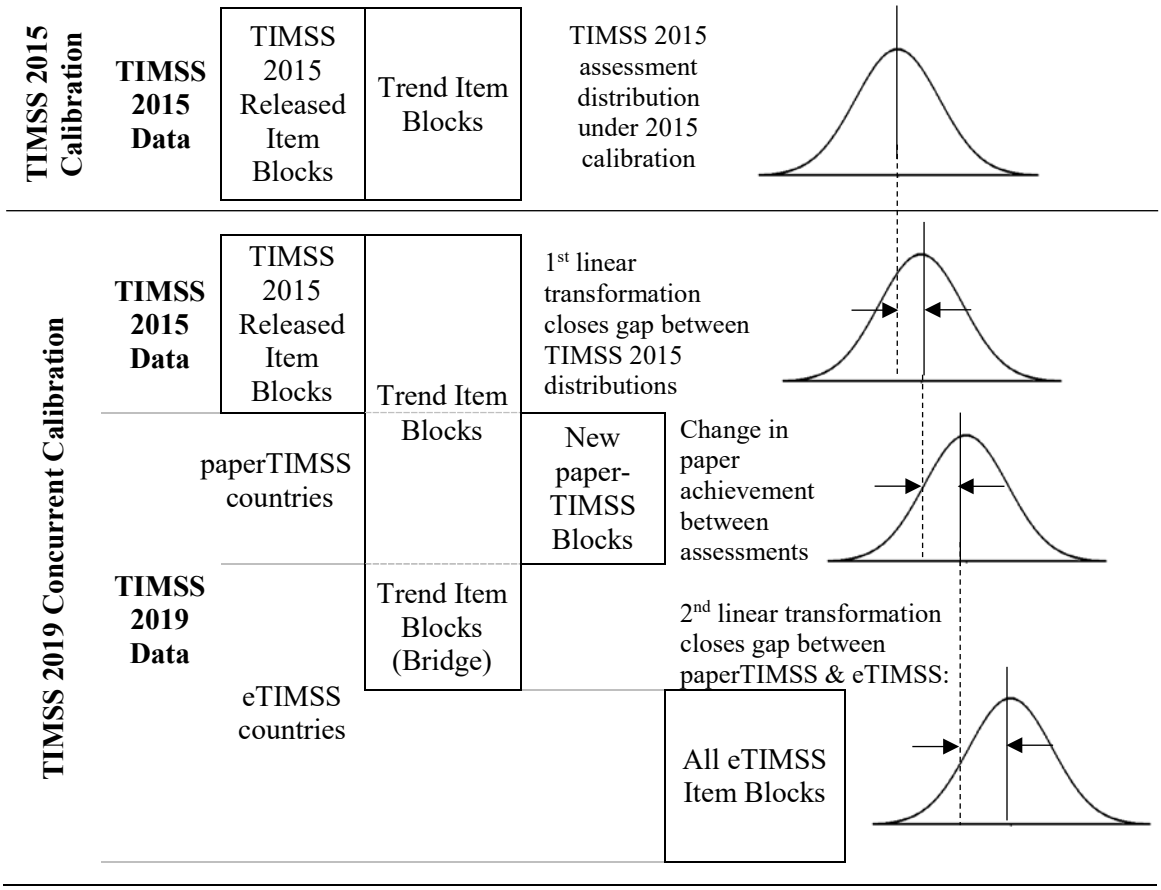
Linking paperTIMSS and eTIMSS Scores for TIMSS 2019

TIMSS will use an equivalent groups, or common population, approach for linking the scales in 2019. This was deemed possible with the large number of trend items that cover a large portion of the TIMSS 2019 item pool in mathematics and science at each grade, as well as the plan for paperTIMSS bridge samples to ensure comparability of TIMSS scale scores for TIMSS 2019.

With the common population linking design, the full eTIMSS sample will be linked with the paperTIMSS bridge samples, with both samples being nationally representative. Rather than assuming equal item parameters for the trend items, unique item parameters will be estimated for all paper and digital items, and the mean proficiency scores will be constrained to be equal across paperTIMSS and eTIMSS to adjust for the differences in psychometric properties of items between modes. This will place paperTIMSS and eTIMSS scores on a common scale.

This approach to link paperTIMSS and eTIMSS scales for TIMSS 2019 is similar to the approach used to link TIMSS 2003 and TIMSS 2007 scales, which was needed due to a change in the booklet design that affected the difficulty of the trend items (Foy, Galia, & Li, 2008). The approach, illustrated in Exhibit 3.26, involves implementing a second linear transformation to that described in the previous section to align the paperTIMSS and eTIMSS score distributions.

Exhibit 3.26: Concurrent Calibration Model for TIMSS 2019



Source: Adapted from Foy & Lin (2016) and Foy, Galia, & Li (2008).

For trend countries transitioning to eTIMSS, two calibrations will be conducted. First, eTIMSS items will be calibrated based on data from all eTIMSS countries. Then, the typical concurrent calibration for paperTIMSS will include TIMSS 2015 data, paperTIMSS 2019 data, and 2019 paper bridge data for eTIMSS countries.

After following the typical concurrent calibration approach and applying a linear transformation to align the distribution of the TIMSS 2015 data with the distribution of the paperTIMSS 2019 data, the second transformation for eTIMSS countries will align the distribution of the eTIMSS scores with the distribution of the of the paperTIMSS

scores. The transformation is the same as specified in equations (3.5), (3.6), and (3.7), but using the already transformed paperTIMSS data and the eTIMSS 2019 data. Then, the eTIMSS 2019 scores will be directly comparable with paperTIMSS 2019 scores, as well as TIMSS scores from all previous assessment cycles.

Conclusion

Despite strong evidence for the comparability of paperTIMSS and eTIMSS scores and the plan in place to ensure the preservation of trend measurements, further analysis is warranted to investigate possible sources of construct irrelevant variance to which the mode effect may be attributed. The complex methodology used for estimating TIMSS achievement scores also warrants further exploration of construct equivalence to better understand the nature of the mode effect, even with a within-subjects design (Buerger et al., 2016). Therefore, score comparability is further examined in relation to external criteria in the next chapter.

Chapter 4: Analysis of the eTIMSS Item Equivalence Database

This chapter includes an investigation of student-level predictors using the eTIMSS Item Equivalence Database to further examine the nature of the differences between paperTIMSS and eTIMSS scores. If student characteristics such as computer familiarity explain any variance in achievement between modes, efforts can be made by TIMSS country participants to provide students experience in using technology in educational contexts in order to mitigate these differences and reduce the presence of construct irrelevant variance in students' scores. In addition, analyses of paperTIMSS scores, eTIMSS scores, and the differences between them by student subgroups can provide additional information about the equivalence of the mathematics and science constructs between modes (Randall, Sireci, Li, & Kaira, 2012). If two scores are comparable, then they should have the same degree of relationship with other related measures (APA, 1986; DePascale et al., 2016; Winter, 2010).

The student questionnaire administered at the fourth and eighth grades as part of the eTIMSS Pilot / Item Equivalence Study included items about demographic characteristics, several proxy items measuring socioeconomic status, and newly developed items asking about students' self-efficacy in using computers and tablets. Analyzing these background variables with respect to paperTIMSS and eTIMSS scores can provide further information about the effect of the mode of administration on TIMSS scale scores and the influence of student characteristics theorized to be associated with

score-level mode effects. The results will also contribute to the existing body of literature about the characteristics of students that contribute to testing mode effects.

Predictors of Testing Mode Effects: A Review of the Literature

Student-level predictors of academic achievement have received renewed interest by researchers in relation to performance on digital assessments. Previous studies have found that the magnitude and direction of score-level mode effects can vary depending on several student-level factors, including:

- Measures of socioeconomic status, including parents' level of education (Bennett et al., 2008) and access to computers and tablets (Jerrim, 2016; MacCann, 2006; Zhang et al., 2016)
- Gender (Cooper, 2006; Gallagher et al., 2002; Jerrim, 2016; Parshall & Kromrey, 1993)
- Attitudes and familiarity with using computers and tablets (Bennett et al., 2008; Chen et al., 2014; Cooper, 2006; Russell, 1999; Sandene et al., 2005; Zhang et al., 2016)

However, research findings can vary by grade level, subject area assessed for the outcome, the measurement properties of predictor variables, and the particular assessment context. Therefore, analysis of these variables using the eTIMSS Item Equivalence Database can help determine their relationships with students' performance on eTIMSS and the degree that the relationships differ from those with paperTIMSS. The following sections provide a comprehensive summary of the mode effect literature for each of the above measures.

Measures of Socioeconomic Status

There are increasing concerns among researchers about a "digital divide," in which students with less access to technology and less experience may be more disadvantaged on computer-based assessments than on paper-based assessments (Bennett et al., 2008; Cooper, 2006; MacCann, 2006; Jerrim, 2016; Zhang et al., 2016). For example, MacCann (2006) found that students of lower socioeconomic status tend to have lower science scores on a computer-based test, even after controlling for paper-based scores. Also, Bennett et al. (2008) found that students who had at least one parent with a college degree performed better on a computer-based mathematics test than a similar group of students who took the same test on paper.

Unfortunately, in many countries, socioeconomically disadvantaged students tend to have less access to technology than socioeconomically advantaged students (OECD, 2015), and more access to digital devices and the Internet both at home and in school are associated with higher levels of technology self-efficacy (Lei & Zhou, 2012). Conversely, students who have less access to technology tend to use computers and tablets less often and may have stronger negative attitudes toward using digital devices for test taking (Cooper, 2006; Lei & Zhou, 2012; Spiezia, 2010; Tømte & Hatlevik, 2011).

Gender

The impact of students' gender on their digital test performance may be of concern for the eTIMSS assessment of mathematics and science. TIMSS 2015 results show that fourth grade boys had higher mathematics achievement than girls in 18 of the 49 countries that participated (Mullis, Martin, Foy, & Hooper, 2016). Also, gender has consistently been found to be associated with students' attitudes toward using technology

(Cooper, 2006; Spiezia, 2010; Tømte & Hatlevik, 2011), which may explain females' relatively lower performance on computer-based assessments (Gallagher et al., 2002; Jerrim, 2016; Parshall & Kromrey, 1993). Males report using technology more than females and have higher levels of confidence in their ability to use technology (Cooper, 2006; Spiezia, 2010; Tømte & Hatlevik, 2011). Male students also tend to be exposed to technology at an earlier age than females (OECD, 2015).

Computer and Tablet Experience

Overall, students with relatively less experience using computers do worse on computer-based tests compared to paper-based tests (Bennett et al., 2008; Russell, 1999; Sandene et al., 2005; Zhang et al., 2016), and students who use technology more often tend to have greater self-efficacy for using digital devices and generally more positive attitudes toward using them in learning contexts (Chen et al., 2014; Tømte & Hatlevik, 2011; Zhong, 2011).

Familiarity with digital devices can serve to advantage students on computer-based assessments. Studies of mode effects on the NAEP mathematics assessments found that students who were more familiar with computers, as measured by input ability and accuracy (e.g., typing and navigating the assessment), had significantly higher scores on the computer-based assessment after controlling for paper scores (Bennett et al., 2008; Sandene et al., 2005). Similar results were found for the NAEP writing assessment (Horkey et al., 2006). Russell (1999) also found that performance differences favoring paper scores over computer-based scores on a mathematics test of constructed response items were moderated by the level of keyboard skills students had. Similarly for tablet

devices, Chen et al. (2014) found that tablet familiarity lessened the negative effects of the digital screen on reading comprehension.

The effect of students' level of familiarity with digital devices on computer-based test scores may be explained through the level of digital self-efficacy, or confidence, that students have as a result of their computer and tablet experiences. The literature suggests that students who are less familiar with digital devices have greater anxiety toward using devices and, in turn, performance worse on digital assessments (Cooper, 2006; Tømte & Hatlevik, 2011). Similarly, students with higher levels of confidence for using technology have higher achievement, as measured by both paper and digital assessments (Luu & Freeman, 2010; Pruet, Ang, & Farzin, 2016; Zhang et al., 2016).

However, frequency of computer use by students is not always a reliable predictor of computer-based assessment performance. Researchers have suggested that negative or "hill-shaped" relationships between the frequency of student computer use and achievement are dependent on the nature of the computer use (Bundsgaard & Gerick, 2017; OECD, 2015). For example, several studies have concluded that too much computer use of the wrong type, such as playing games and using social media, is negatively associated with achievement, regardless of the mode in which achievement was measured (Bundsgaard & Gerick, 2017; Kubiatko & Vlkova, 2010; O'Dwyer, Russell, Bebell, & Tucker-Seeley, 2005; OECD, 2015; Skyrabin et al., 2015; Spiezia, 2010; Tømte & Hatlevik, 2011). Similarly, more frequent computer use for school-related purposes may be reported by students with lower achievement, such as students who use computers for remedial schoolwork (O'Dwyer et al., 2005; Skyrabin et al., 2015; Zhang

et al., 2016). Therefore, greater familiarity with technology is not necessarily sufficient to foster higher achievement on computer-based assessments.

The positive effects of technology experience on test performance may be enhanced if there is considerable experience in using technology for educational reasons (Kubiatko & Vlkova, 2010). Researchers suggest that familiarizing students with using computers and tablets in the classroom may help mitigate performance differences on paper versus digital assessments (Davis et al., 2017; Strain-Seymour et al., 2013).

Research Summary

Overall, the research suggests that students' experiences with technology vary by students' socioeconomic status and gender, and the extent and nature of their experiences are predictive of their self-efficacy for using computers and tablets. In turn, greater confidence for using technology is associated with higher digital test performance. However, familiarity with computers alone is not always predictive of achievement.

As evidenced by the review of literature, findings across studies may be related to the different ways computer familiarity and self-efficacy are measured, rapid changes in exposure to technology, and increasing improvements to digital assessment systems (Kingston, 2008; McDonald, 2002; Way et al., 2016). Examining these relationships in the TIMSS international context may provide new insight into the impact that computer and tablet self-efficacy, or "digital self-efficacy," has on digital test performance beyond that of self-efficacy measures on paper-and-pencil test performance.

Analysis Methodology

Analysis of the eTIMSS Item Equivalence Database was conducted separately for mathematics and science at the fourth and eighth grades to address the following research questions:

1. Does the TIMSS mode of administration differentially affect subgroups of students based on gender?
2. Does the magnitude of mode-related performance differences vary by measures of socioeconomic status?
3. Does digital self-efficacy have similar relationships with achievement measured by paperTIMSS and eTIMSS?

Addressing these questions involved examining whether characteristics of students explain any of the variance between paperTIMSS and eTIMSS scores. The results are of interest for examining scale score comparability, through exploring the presence of construct irrelevant variance in the mode effects. The results can also help identify groups of students who may be at risk for exhibiting mode effects on TIMSS.

Description of the eTIMSS Item Equivalence Database

The eTIMSS Item Equivalence Database contains data collected from 21 countries at the fourth grade and 11 countries at the eighth grade that participated in the eTIMSS Pilot / Item Equivalence Study. At each grade, the database includes mathematics and science achievement scores for each student based on paperTIMSS and eTIMSS, respectively. These achievement scores are in the form of five plausible values. Conducting analysis across all five plausible values accounts for the variation among the estimates that result from the estimation procedure.

Data are included about students' gender, socioeconomic status, and self-efficacy for using computers and tablets. The background data were collected through a short student questionnaire given at the end of the eTIMSS test session on a computer or tablet. Information about students' gender and the digital device used for eTIMSS delivery—PC or tablet—was collected through Student Tracking Forms (see Chapter 2).

Exhibit 4.1 presents the total unweighted sample sizes and percentages of students by gender at the fourth and eighth grades. Across countries, there is approximately equal distribution of girls and boys—with 51 percent girls and 49 percent boys at the fourth grade, and 53 percent girls and 47 percent boys at the eighth grade.

Exhibit 4.1: eTIMSS Item Equivalence Database—Percentage of Students by Gender (Unweighted)

Grade	Total Cases	Sex of Student		
		Valid N	Percent Girls	Percent Boys
Fourth Grade (21 countries)	16,894	16,769	51%	49%
Eighth Grade (11 countries)	9,164	9,102	53%	47%

Exhibit 4.2 presents the percentages of students who used PCs versus tablets for taking the eTIMSS Pilot / Item Equivalence Study at the fourth and eighth grades. Overall, most countries used either all PCs or all tablets for eTIMSS, with most countries using PCs. Compared to the fourth grade, where 66 percent of students used PCs and 33 percent used tablets, a larger proportion of eighth grade students used PCs compared to tablets across countries (85% vs. 15%).

Exhibit 4.2: eTIMSS Item Equivalence Database—Percentage of Students by eTIMSS Device (Unweighted)

Grade	Total Cases	eTIMSS Device		
		Valid N	Percent PC	Percent Tablet
Fourth Grade (21 countries)	16,894	16,873	66%	33%
Eighth Grade (11 countries)	9,164	9,141	85%	15%

Variables of Interest

Based on the literature and the goal of exploring the comparability of paperTIMSS and eTIMSS scores and investigating factors contributing to the mathematics and science score mode effects, three variables are of interest for analysis: gender, socioeconomic status, and self-efficacy for using computers and tablets, or "digital self-efficacy."

Gender

Gender data for each student were collected from participating schools via the Student Tracking Forms (see Chapter 2), as well as from students via the student eTIMSS questionnaire. Following data cleaning procedures from TIMSS 2015 (Meyer, Cockle, & Tavena, 2016), data were imputed from the tracking form data for any missing values in the questionnaire variables. This variable is coded in the database as 1 = "Girl" and 2 = "Boy."

Socioeconomic Status

Proxy measures of students' socioeconomic status are of interest, specifically the "Books in the Home" variable, which historically has shown to be a strong predictor of

achievement in TIMSS (e.g., Mullis, Martin, Foy, & Hooper, 2016; Mullis, Martin, & Hooper, 2017). In addition, eighth grade students provided the level of education of their parents. Exhibit 4.3 includes descriptions of these variables.

Exhibit 4.3: eTIMSS Pilot / Item Equivalence Study eTIMSS Questionnaire Items Measuring Socioeconomic Status

1. Books in the Home—Fourth and Eighth grades

About how many books are there in your home? (Do not count magazines, newspapers, or your school books.)

1 = None or very few (0-10 books)
2 = Enough to fill one shelf (11-25 books)
3 = Enough to fill one bookcase (26-100 books)
4 = Enough to fill two bookcases (101-200 books)
5 = Enough to fill three or more bookcases (more than 200)

2. Parents' Education—Eighth grade

A. What is the highest level of education completed by your mother (or stepmother or female guardian)?

B. What is the highest level of education completed by your father (or stepfather or male guardian)?

1 = Some <Primary education—ISCED Level 1 or Lower secondary education—ISCED Level 2> or did not go to school
2 = <Lower secondary education—ISCED Level 2>
3 = <Upper secondary education—ISCED Level 3>
4 = <Post-secondary, non-tertiary education—ISCED Level 4> or <Short-cycle tertiary education—ISCED Level 5>
5 = <Bachelor's or equivalent level—ISCED Level 6>
6 = <Postgraduate degree: Master's—ISCED Level 7 or Doctor—ISCED Level 8>
7 = I don't know

Source: eTIMSS Pilot / Item Equivalence Study eTIMSS Questionnaire, Fourth and Eighth Grades
Note: Items are confidential. Do not cite or circulate.
Brackets < > around answer options indicate that countries adapt this item according to the structure of the education system in the country.

A derived variable for "Parents' Education" was created using the two items—mother's highest level of education and father's highest level of education. Each item was

recoded so that 5 = “Finished University or Higher” (original categories 5 and 6); 4 = “Finished Post-Secondary Education” (original category 4); 3 = “Finished Upper Secondary” (original category 3); 2 = “Finished Lower Secondary” (original category 2); 1 = “Finished Some Primary or Lower Secondary or Did Not Go to School” (original category 1); and 0 = “Not applicable” (original category 7). Using these categories, the smaller value of the recoded variables became the final value for the level of Parents’ Education. Category 0 (“Not applicable”) was set to missing.

Digital Self-Efficacy

A one-parameter IRT (Rasch) scale of digital self-efficacy was constructed using data from the two groups of questionnaire items described in Exhibit 4.4. Appendix A provides the details of the analyses conducted in constructing the scale.

The IRT scale measuring students’ digital self-efficacy was constructed following the general procedures used for TIMSS and PIRLS background scales (Martin, Mullis, Hooper, Yin, Foy, & Fishbein, 2017; Martin, Mullis, Hooper, Yin, Foy, & Palazzo, 2016). After selecting the scale items and ensuring the items constitute a unidimensional and reliable scale, a Rasch partial credit IRT model (Masters, 1982) was used to construct the scale with ACER’s Conquest software (Adams, Wu, & Wilson, 2015). The models converged at both grades, and all items showed good fit to the model (infit < 1.3). As an added assurance of scale quality, concerns of autocorrelation were addressed using Winsteps 4.0.0 software (Linacre, 2017), which allowed for closely examining the person residuals that resulted from the model. Winsteps resulted in approximately the same item and person estimates as Conquest.

Exhibit 4.4: eTIMSS Pilot / Item Equivalence Study eTIMSS Questionnaire Items Measuring Digital Self-Efficacy

1. How much do you agree with these statements?

1 = Agree a lot
2 = Agree a little
3 = Disagree a little
4 = Disagree a lot

a) I am good at using a computer

b) I am good at typing

c) It is easy for me to find information on the Internet

2. Can you do each of the following?

1 = I definitely can
2 = I probably can
3 = I probably cannot
4 = I definitely cannot

a) Write sentences and paragraphs using a computer

b) Use a touchscreen on a computer, tablet, or smartphone

c) Type using the correct fingers

d) Draw a picture using a computer

e) Find information on the Internet

Source: eTIMSS Pilot / Item Equivalence Study eTIMSS Questionnaire, Fourth and Eighth Grades
Note: Items are confidential. Do not cite or circulate.

Scale scores for each student were produced using weighted maximum likelihood estimation (Warm, 1989). Using a linear transformation, the scale was transformed onto the TIMSS context questionnaire scale reporting metric, with a mean of 10 and a standard deviation of 2. The scale was validated for analysis to confirm that the scale has at least a small, positive relationship with achievement. Then, a benchmarking procedure was used to classify students' scores into meaningful "Low," "Medium," and "High" categories of digital self-efficacy (see Appendix A).

Analysis Procedures

The analysis included two phases for each grade and subject to address the research questions. In Phase 1, a descriptive analysis of mathematics and science mean scale scores by student subgroups according to the three variables of interest was conducted to explore whether difference scores between paperTIMSS and eTIMSS were approximately the same. Then, an analysis of variance (ANOVA) was conducted to explore whether the student-level predictor variables had a significant interaction with mode in predicting variance between paperTIMSS and eTIMSS scores. Phase 2 included a multiple linear regression analysis of the difference scores between paperTIMSS and eTIMSS to examine the percentage of variance that the predictor variables accounted for in the models. The results were compared to the ANOVA results to explore inconsistencies.

Phase 1 Analysis Procedures

Relationships between the variables of interest and achievement were examined by grade with mean scores for mathematics and science scores, respectively. The IEA IDB Analyzer script for SPSS was used to compute international average achievement scores and standard errors for paperTIMSS, eTIMSS, and their differences by category of each variable of interest: socioeconomic status (Books in the Home for both grades and Parents' Education for eighth grade students), gender, and digital self-efficacy. The IEA script computes the mean of each plausible value, aggregates the results, and computes jackknife standard errors using the balanced repeated replication method (Foy & LaRoche, 2016; Rust, 2014), resulting in estimates that account for both sampling error and measurement error in the achievement scores. While not accurate of population data

without random samples, the standard errors are more realistic than those based on a simple random sampling formula.

The IEA scripts also computed the international average percentage of students in each variable category. The SENWGT was applied in each analysis so that each country contributed equally to the estimates. Using the resulting estimates of the average mode effects by subgroup, plots of the mean difference scores with 95 percent confidence intervals were created to accompany the numerical results. To facilitate comparison across grades and subjects, plots for mathematics and science were combined at each grade.

The repeated measures experimental design of the eTIMSS Pilot / Item Equivalence Study calls for the use of a mixed design ANOVA model to test whether the magnitude of the mode effects between paperTIMSS and eTIMSS varied significantly by student subgroups. A $5 \times 2 \times 3 \times 2$ mixed design ANOVA model was tested for each grade and subject using SPSS. The student-level factors were tested for their within-subject interactions with mode in predicting achievement: socioeconomic status as measured by Books in the Home (5 levels), gender (2 levels), and digital self-efficacy (3 levels). Three- and four-way interaction effects among the predictors were also included in the full factorial models. Following proper secondary analysis procedures for using plausible values, each model was run in SPSS five times—once for each pair of plausible values as the within-subjects variables (paperTIMSS and eTIMSS)—and results aggregated for interpretation. This accounted for the variation among the achievement estimates in the results of the analysis. Cases were weighted using SENWGT so that each country contributed equally to the results.

Assumptions of mixed ANOVA models were assessed for each model using methods that accommodate large sample sizes. Normality of the dependent variables were assessed by examining histograms and normal Q-Q plots of paperTIMSS scores and eTIMSS scores by category of each predictor variable. Heteroscedasticity was assessed with Hartley's F_{max} variance ratio (Pearson & Hartley, 1954). Sphericity was assessed using Mauchly's test. Residual plots were also examined.

The within-subjects model results are of interest to determine which variables interact with mode to predict achievement. For each within-subjects effect, partial eta squared (η_p^2) is reported as a measure of effect size, with benchmark values of 0.14 for large effects, 0.06 for medium effects, and 0.01 for small effects (Kirk, 1996). These can be interpreted as the proportion of variance that the variable (or variable interaction) accounts for between paperTIMSS and eTIMSS scores. Due to the uneven sample sizes, particularly for the categorical digital self-efficacy variable, three- and four-way interaction effects are only reported in the next section if they were statistically significant in the model. Pairwise comparisons based on estimated marginal means were adjusted for multiple comparisons with a Bonferroni adjustment.

Phase 2 Analysis Procedures

As a second method of analyzing the influence of the predictor variables on the mode effects and to more accurately estimate the percentage of variance accounted for by the predictor variables in the mode effects, a multiple linear regression analysis was conducted in SPSS using the IEA IDB Analyzer script for linear regression. Like the mean score analysis, this script uses the jackknife repeated replication method to compute sampling error (Foy & LaRoche, 2016; Rust, 2014), runs each regression model five

times, and aggregates the results for interpretation. Although standard errors are not accurate estimates of population data, the jackknifing provided more realistic estimates than the ANOVA models.

The multiple linear regression model was specified for each grade and subject so that each country contributed equally to the analysis with the senate weight (SENWGT). The outcome variable was the set of plausible values for the difference scores between paperTIMSS and eTIMSS (PVDIFF), which were computed in SPSS using each respective pair of paperTIMSS and eTIMSS plausible values. The following model was specified for each grade and subject:

$$(PVDIFF)_{ij} = B_0 + B_1(DSE)_{ij} + B_2(SES_1)_{ij} + \ldots + B_5(SES_4)_{ij} + B_6(Gender)_{ij} + \varepsilon_{ij} \qquad (4.1)$$

In this model, the continuous digital self-efficacy scale variable (DSE) was used instead of the categorical variable used in the ANOVA. Books in the Home was entered as a dummy-coded predictor variable for socioeconomic status where "0-10 books" was the reference category (0) and 1 = "11-25 books" (SES_1); 2 = "26-100 books" (SES_2); 3 = "101-200 books" (SES_3); and 4 = "More than 200 books" (SES_4). Gender was entered as dummy-coded predictor variable where "Girls" were the reference group (0) and 1 = "Boys." Examination of VIF, tolerance, and condition index values indicated no multicolinearity among predictor variables in all four models.

Unstandardized regression coefficients (*B*) for each of the predictors were interpreted and examined for statistical significance, and R^2 values for the entire model were examined to determine the percentage of variance in the difference scores accounted for by the predictor variables.

Results

The results of scale score analysis for the eTIMSS Pilot / Item Equivalence Study presented in Chapter 3 revealed mode effects on the TIMSS mathematics and science scores. The mode effects were larger for mathematics, with international average differences between paperTIMSS and eTIMSS scores of 14 points at the fourth and eighth grades. The mode effects for science were smaller—8 points at the fourth grade and 7 points at the eighth grade, on average. The results of the analysis of the eTIMSS Item Equivalence Database by subgroups helped to determine whether student characteristics explain these differences.

Phase 1 Results

Three sub-sections below present the results of the analysis for Phase 1—one for each predictor variable of interest: socioeconomic status, gender, and digital self-efficacy. Interpretation of mean scores by subgroup is included for each grade and subject, with statistical significance of the within-subjects effects and the approximate measure of effect size provided based on the ANOVA model. Each section includes plots of international average mode effects with 95 percent confidence intervals to illustrate the relationships.

The ANOVA models detected one significant three-way interaction effect on fourth grade mathematics scores among mode, socioeconomic status, and digital self-efficacy ($\eta_p^2 = 0.001$). However, the effect accounted for less than 1 percent of the variance within-subjects overall. Closer examination of the relationships revealed that differences were negligible after accounting for differences in sample size among the marginal means. Therefore, the following sections only include analysis of main effects.

Socioeconomic Status

Exhibit 4.5 presents mean paperTIMSS, eTIMSS, and difference scores by Books in the Home for the fourth grade. The numbers in the far right "Difference" column of the exhibit for mathematics show that the size of the difference between paperTIMSS and eTIMSS scores was slightly higher for students in the three high categories of socioeconomic status. Fourth grade students in the "0-10 books" and "11-25 books" categories had average differences of 12 and 13 points in mathematics, respectively, while students in the "26-100 books," "101-200 books," and "More than 200 books" categories had mean differences of 15, 16, and 15 points, respectively.

Exhibit 4.5: International Average Scale Scores by Books in the Home—Fourth Grade (21 countries)

Books in the Home	Valid Cases	Average Percent of Students	International Average Scale Scores		
			paperTIMSS	eTIMSS	Difference
Mathematics					
0-10 books	1,947	12 (0.4)	481 (2.3)	469 (2.5)	12 (1.4)
11-25 books	4,180	26 (0.5)	512 (1.7)	499 (1.6)	13 (0.8)
26-100 books	5,411	33 (0.4)	538 (1.6)	523 (1.7)	15 (0.7)
101-200 books	2,811	17 (0.4)	547 (2.2)	531 (2.2)	16 (0.9)
More than 200 books	2,201	13 (0.3)	542 (2.4)	527 (2.4)	15 (1.1)
Science					
0-10 books	1,947	12 (0.4)	480 (2.5)	469 (2.6)	10 (1.7)
11-25 books	4,180	26 (0.5)	510 (1.7)	502 (1.7)	7 (0.9)
26-100 books	5,411	33 (0.4)	537 (1.8)	528 (1.8)	9 (0.7)
101-200 books	2,811	17 (0.4)	544 (2.4)	537 (2.3)	7 (0.9)
More than 200 books	2,201	13 (0.3)	543 (2.5)	535 (2.5)	8 (1.0)

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

The ANOVA results showed that effect of socioeconomic status on the mathematics mode effect was small, but significant at the fourth grade, $F_{M4}(4, 15018) = 6.22$, $p < 0.001$, $\eta_p^2 = 0.002$. This effect size indicates that

socioeconomic status $\times$ mode accounted for less than 1 percent of the variance within-subjects overall. After pooling standard errors and applying a Bonferroni correction, results indicated that the strongest effects of socioeconomic status occurred for students with the most books in the home—the "26-100 books" category and the "101-200 books" category.

In science, students in the lowest socioeconomic status category exhibited the largest mode effects (10-point average difference), with slightly smaller average differences for the two highest categories (7 and 8 points, respectively). The effect for science was not significant, $F_{S4}(4, 15018) = 1.117$, $p > 0.05$, $\eta_p^2 < 0.001$.

Exhibit 4.6 shows the mean scores by Books in the Home for the eighth grade. The results for mathematics were similar to the fourth grade, where higher score differences were seen for the two highest categories, with 15- and 16-point average differences between paperTIMSS and eTIMSS, respectively. In science, the largest score differences were exhibited by students in the lowest "0-10 books" category and the highest "More than 200 books" (9 points, on average), and the smallest difference was for the "101-200 books" category (4 points). The effect of socioeconomic status was significant for mathematics, but not for science, $F_{M8}(4, 6849) = 2.70$, $p < 0.05$, $\eta_p^2 = 0.002$, $F_{S8}(4, 6849) = 1.43$, $p > 0.05$, $\eta_p^2 = 0.001$. Both effects were very small, accounting for less than 1 percent of the variance between modes overall.

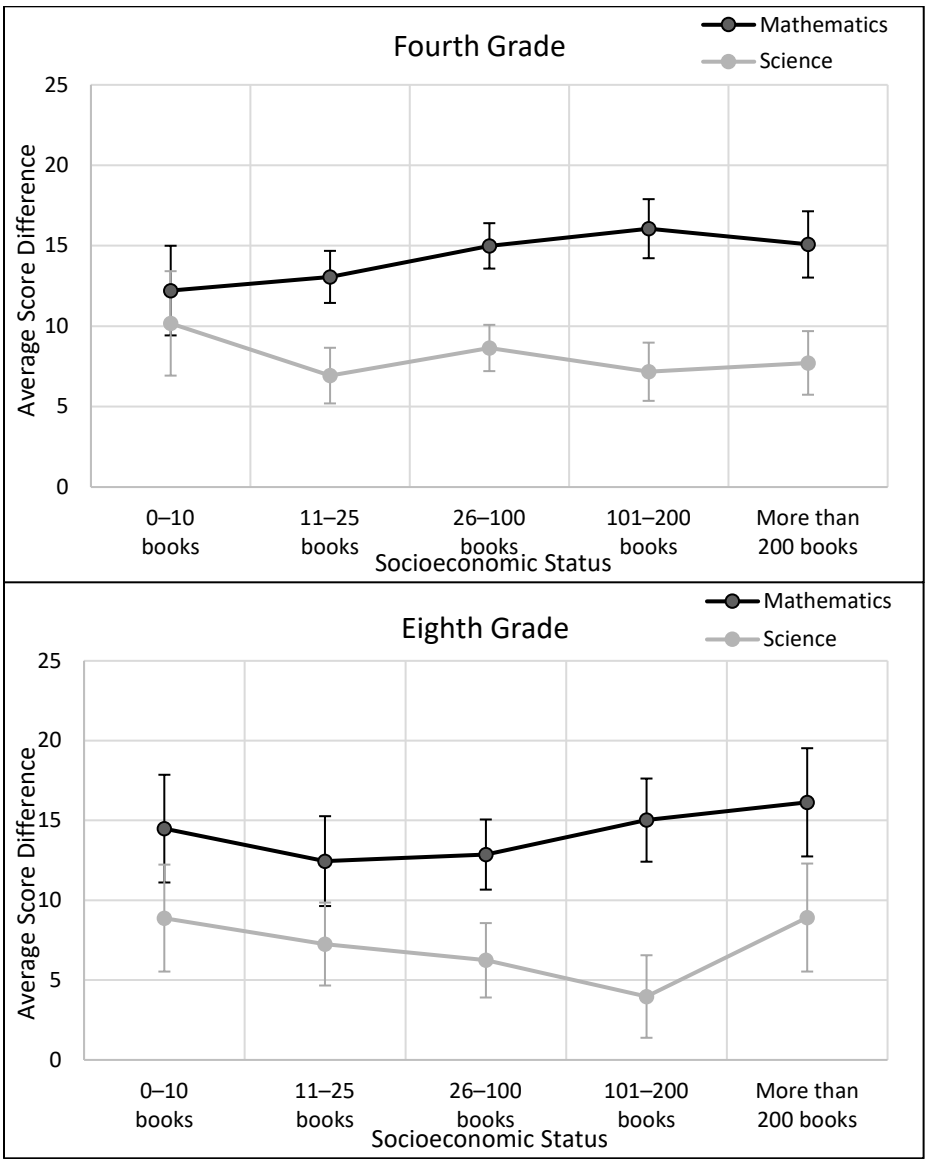
Exhibit 4.6: International Average Scale Scores by Books in the Home—Eighth Grade (11 countries)

Books in the Home	Valid Cases	Average Percent of Students	International Average Scale Scores		
			paperTIMSS	eTIMSS	Difference
Mathematics					
0-10 books	1,241	14 (0.5)	468 (3.3)	453 (3.4)	14 (1.7)
11-25 books	2,160	25 (0.6)	490 (3.0)	477 (2.9)	12 (1.4)
26-100 books	2,670	30 (0.5)	517 (2.4)	505 (2.4)	13 (1.1)
101-200 books	1,450	16 (0.5)	540 (3.4)	525 (3.6)	15 (1.3)
More than 200 books	1,454	16 (0.6)	543 (3.2)	526 (3.3)	16 (1.7)
Science					
0-10 books	1,241	14 (0.5)	475 (3.5)	467 (3.4)	9 (1.7)
11-25 books	2,160	25 (0.6)	498 (3.1)	491 (2.9)	7 (1.3)
26-100 books	2,670	30 (0.5)	528 (2.7)	522 (2.6)	6 (1.2)
101-200 books	1,450	16 (0.5)	549 (3.7)	545 (3.7)	4 (1.3)
More than 200 books	1,454	16 (0.6)	553 (3.7)	545 (3.5)	9 (1.7)

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

Exhibit 4.7 shows plots of the mean difference scores by socioeconomic status for both grades, which illustrate an overall small, positive relationship between Books in the Home and the magnitude of the mathematics score differences. Science difference scores appear to fluctuate more randomly by Books in the Home category at both grades. Within the margin of error, the variation among the difference scores are negligible.

Exhibit 4.7: Average Mathematics and Science Mode Effects by Socioeconomic Status



Error bars reflect 95% confidence intervals of the estimated mean score differences.

The impact of socioeconomic status on the mathematics and science score mode effects was further examined at the eighth grade using the derived Parents' Education variable. Exhibit 4.8 presents international average scale scores by level of parents' education for eighth grade students. There was no clear influence of parents' education on students' performance differences between paperTIMSS and eTIMSS. In

mathematics, the highest average differences between paperTIMSS and eTIMSS scores were seen for students whose parents completed post-secondary education (15 points) and lower secondary education (16 points) and the lowest average differences were seen for students whose parents completed lower secondary education (8 points).

Exhibit 4.8: International Average Scale Scores by Parents' Highest Level of Education—Eighth Grade (11 countries)

Parents' Education	Valid Cases	Average Percent of Students	International Average Scale Scores					
			paperTIMSS		eTIMSS		Difference	
Mathematics								
Some Lower Secondary or Less*	152	2 (0.2)	458	(13.9)	442	(15.0)	16	(7.5)
Lower Secondary	425	6 (0.3)	463	(5.8)	455	(7.6)	8	(3.3)
Upper Secondary	1,547	21 (0.6)	487	(3.2)	474	(3.3)	13	(1.7)
Post-Secondary	1,256	18 (0.6)	512	(3.3)	496	(3.4)	15	(1.7)
University or Higher	3,858	54 (1.0)	532	(2.4)	518	(2.4)	14	(1.1)
Science								
Some Lower Secondary or Less*	152	2 (0.2)	461	(12.1)	455	(12.3)	7	(4.2)
Lower Secondary	425	6 (0.3)	472	(6.2)	466	(7.2)	6	(2.8)
Upper Secondary	1,547	21 (0.6)	499	(3.5)	491	(3.6)	7	(1.4)
Post-Secondary	1,256	18 (0.6)	522	(3.7)	512	(3.6)	10	(1.7)
University or Higher	3,858	54 (1.0)	544	(2.8)	537	(2.6)	7	(1.2)

( ) Standard errors appear in parentheses. Because of rounding some results may appear inconsistent.
* Contains less than 3 percent of cases. Interpret scale scores with caution.

Similar results were found for science, where students with parents who completed post-secondary education had paperTIMSS scores 10 points higher than eTIMSS scores, on average. Students whose parents completed lower secondary education showed the lowest mode effects, on average, with a difference of only 6 score points. However, the results for the two lowest categories should be interpreted carefully because of the small samples sizes.

Gender

Exhibit 4.9 presents the international average paperTIMSS, eTIMSS, and difference scores for girls and for boys at the fourth grade. At the fourth grade, the score difference was two points higher for girls than for boys in mathematics, on average (15 points vs. 13 points), but the associated effect was negligible in size and non-significant in the ANOVA model, $F_{M4}(1, 15018) = 0.25$, $p > 0.05$, $\eta_p^2 < 0.001$.

In science, the score difference was two points *lower* for girls than for boys (7 points vs. 9 points). This effect of gender on the fourth grade science mode effect was statistically significant, $F_{S4}(1, 15018) = 5.77$, $p < 0.05$, $\eta_p^2 < 0.001$. However, the effect size is very small and examining the paperTIMSS and eTIMSS scale scores show negligible average performance differences between boys and girls overall.

Exhibit 4.9: International Average Scale Scores by Gender—Fourth Grade (21 countries)

	Valid Cases	Average Percent of Students	International Average Scale Scores		
			paperTIMSS	eTIMSS	Difference
Mathematics					
Girls	8,627	52 (3.0)	524 (1.6)	509 (1.6)	15 (0.8)
Boys	8,142	49 (3.0)	532 (1.8)	518 (1.8)	13 (0.8)
Science					
Girls	8,627	52 (3.0)	526 (1.7)	519 (1.7)	7 (0.6)
Boys	8,142	49 (3.0)	526 (2.1)	517 (2.0)	9 (0.8)

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

Exhibit 4.10 presents the Phase 1 results for gender at the eighth grade. In mathematics, girls had a 14-point mean difference between paperTIMSS and eTIMSS scores, and the results for boys were about the same (13 points). In contrast to the fourth grade results, the mode effect on science scores was somewhat larger for girls than for

boys, on average (9 points vs. 6 points). The ANOVA models found no significant effects of gender on the mathematics and science score differences between paperTIMSS and eTIMSS at the eighth grade, $F_{M8}(1, 6849) = 1.23$, $p > 0.05$, $\eta_p^2 < 0.001$, $F_{S8}(1, 6849 = 2.63$, $p > 0.05$, $\eta_p^2 < 0.001$.

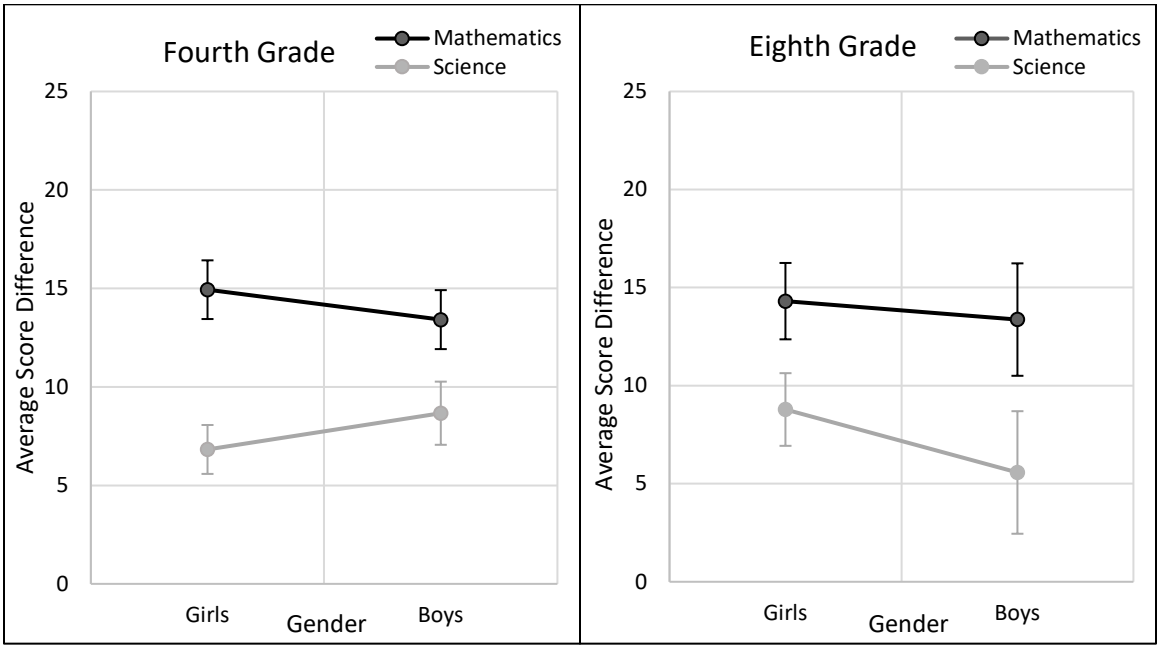
Exhibit 4.10: International Average Scale Scores by Gender—Eighth Grade (11 countries)

	Valid Cases	Average Percent of Students	International Average Scale Scores		
			paperTIMSS	eTIMSS	Difference
Mathematics					
Girls	4,794	53 (1.4)	512 (2.4)	498 (2.5)	14 (1.0)
Boys	4,308	48 (1.4)	511 (3.4)	497 (3.2)	13 (1.5)
Science					
Girls	4,794	53 (1.4)	527 (3.1)	518 (3.0)	9 (0.9)
Boys	4,308	48 (1.4)	516 (3.9)	510 (3.5)	6 (1.6)

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

Exhibit 4.11 presents a graphical display of the estimated mean difference scores by gender. The plots show that the mode-gender relationship at the fourth grade is different for science compared to mathematics, which appears to have approximately equal score differences for girls and boys within the confidence interval. Boys show somewhat larger mode effects compared to girls in science, on average. The plot of means at the eighth grade shows a small advantage for boys compared to girls in both subjects with regard to mode effects, but the differences are negligible within the margin of error.

Exhibit 4.11: Average Mathematics and Science Mode Effects by Gender



Error bars reflect 95% confidence intervals of the estimated mean score differences.

Digital Self-Efficacy

The analysis results suggest that students' level of digital self-efficacy had *some* positive impact on their paperTIMSS-eTIMSS performance discrepancy. However, the effect was non-significant in all four ANOVA models, and uneven sample sizes among the three groups limit the interpretation of the model results.

Exhibit 4.12 shows the international average scores for fourth grade students in mathematics and science by level of digital self-efficacy. Examining eTIMSS and paperTIMSS scores shows a positive relationship between digital self-efficacy and achievement in each mode, respectively. However, there was virtually no effect of digital self-efficacy on fourth grade mathematics score differences, $F_{M4}(2, 15018) = 1.07$, $p > 0.05$, $\eta_p^2 < 0.001$. The "Low" category had an international average difference of 17

points between paperTIMSS and eTIMSS scores, and students in the "Medium" and "High" categories each had 14-point average differences.

In science, students in the "Low" digital self-efficacy category had the highest difference between paperTIMSS and eTIMSS scores, on average, with a difference of 11 score points, compared to 7- and 8-point differences for the "Medium" and "High" categories, respectively. However, the effect was also non-significant for science, $F_{S4}(2, 15018) = 1.62, p > 0.05, \eta_p^2 < 0.001$.

Exhibit 4.12: International Average Scale Scores by Level of Digital Self-Efficacy—Fourth Grade (21 countries)

Digital Self-Efficacy	Valid Cases	Average Percent of Students	International Average Scale Scores		
			paperTIMSS	eTIMSS	Difference
Mathematics					
Low	1,094	7 (0.3)	499 (3.0)	482 (3.0)	17 (1.7)
Medium	4,926	30 (0.4)	520 (1.8)	506 (1.7)	14 (0.9)
High	10,455	63 (0.6)	535 (1.5)	521 (1.5)	14 (0.7)
Science					
Low	1,094	7 (0.3)	494 (3.0)	484 (3.0)	11 (2.0)
Medium	4,926	30 (0.4)	517 (1.8)	510 (1.8)	7 (0.9)
High	10,455	63 (0.6)	535 (1.6)	527 (1.5)	8 (0.6)

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

Exhibit 4.13 presents the mean scores for digital self-efficacy at the eighth grade, which show even less variation among difference scores for each of the three categories of digital self-efficacy. In mathematics, students in the "Low," "Medium," and "High" categories had mean difference scores of 15, 15, and 13 points respectively. The results were similar for science, with mean difference scores of 8, 8, and 6 points respectively. The effects were non-significant in both mathematics and science, $F_{M8}(2, 6849) = 1.95, p > 0.05, \eta_p^2 = 0.001, F_{S8}(2, 6849) = 1.74, p > 0.05, \eta_p^2 < 0.001$.

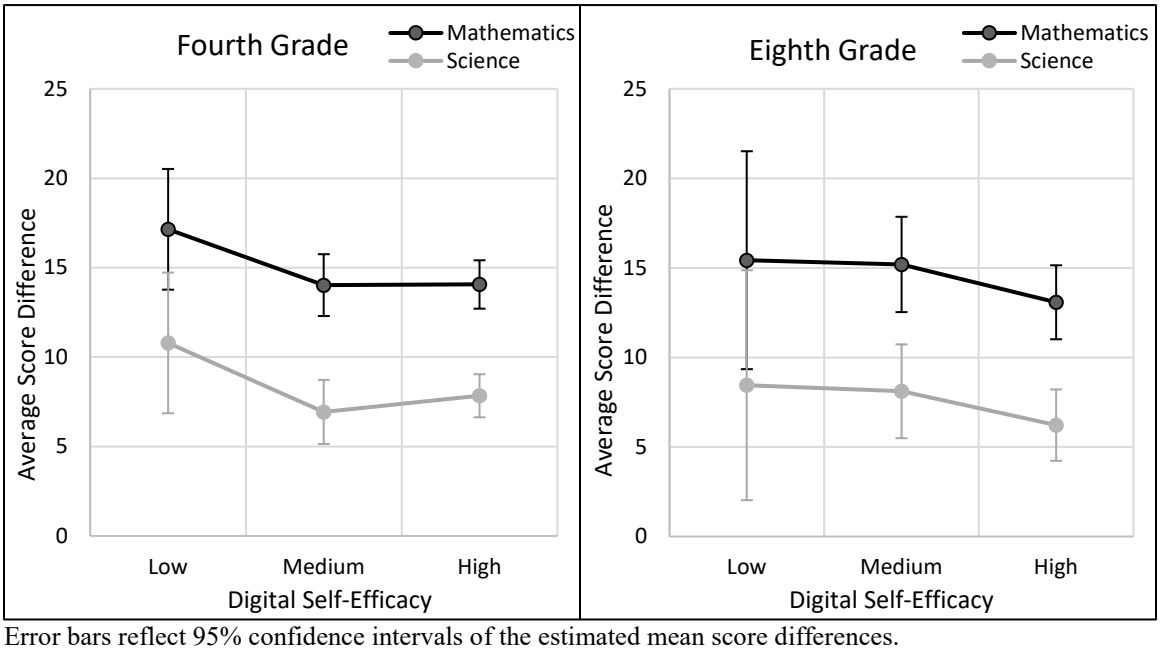
Exhibit 4.13: International Average Scale Scores by Level of Digital Self-Efficacy—Eighth Grade (11 countries)

Digital Self-Efficacy	Valid Cases	Average Percent of Students	International Average Scale Scores		
			paperTIMSS	eTIMSS	Difference
Mathematics					
Low	379	4 (0.3)	462 (7.1)	446 (6.7)	15 (3.1)
Medium	2,302	26 (0.5)	496 (2.9)	481 (3.0)	15 (1.4)
High	6,286	70 (0.6)	519 (2.3)	506 (2.3)	13 (1.1)
Science					
Low	379	4 (0.3)	466 (7.4)	457 (7.1)	8 (3.3)
Medium	2,302	26 (0.5)	505 (3.2)	497 (3.1)	8 (1.3)
High	6,286	70 (0.6)	530 (2.7)	524 (2.6)	6 (1.0)

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

Exhibit 4.14 displays the graphical relationships between the score differences and digital self-efficacy at both grades. The plots illustrate overall small, negative relationships between students' level of digital self-efficacy and size of the score mode effects. Within the margin of error, the differences are negligible.

Exhibit 4.14: Average Mathematics and Science Mode Effects by Digital Self-Efficacy



Error bars reflect 95% confidence intervals of the estimated mean score differences.

Phase 2 Results

In addition to corroborating some of the Phase 1 results, the results of Phase 2 suggest that after controlling for students' digital self-efficacy and gender, socioeconomic status also has a significant influence on science mode effects. However, the predictor variables explain a very small percentage of variance in the mode effects, overall. At the fourth grade, the predictor variables accounted for approximately 2 percent of the variance in mathematics score differences ($R^2 = 0.015$) and 2 percent of the variance in science score differences ($R^2 = 0.016$). At the eighth grade, the predictor variables accounted for approximately 1 percent of the variance in mathematics differences scores ($R^2 = 0.013$) and approximately 2 percent of the variance in science differences scores ($R^2 = 0.016$).

Exhibit 4.15 presents the regression coefficients for the fourth grade models, which are similar to the ANOVA results for mathematics. For mathematics, being in the "26-100 books" category compared to the "0-10 books" reference category had a significant positive association with mode effects, after controlling for students' digital self-efficacy and gender. Having more books at home was associated with larger performance differences between paperTIMSS and eTIMSS ($\beta_3 = 2.87$, $p < 0.05$). Similarly, having "101-200 books" at home was associated with larger mode effects ($\beta_4 = 3.99$, $p < 0.01$). The effect for students with the most books, in the "More than 200 books" category, was similar, but non-significant. Also consistent with the Phase 1 results, being a boy was associated with smaller mode effects in mathematics, but significantly larger mode effects in science ($\beta_6 = 1.91$, $p < 0.05$).

Exhibit 4.15: International Average Regression Coefficients—Fourth Grade (21 countries)

	Mathematics		Science	
	B	***SE***	***B***	***SE***
Intercept	14.73**	2.53	10.25**	2.96
1. Digital Self-Efficacy	-0.21	0.19	-0.12	0.20
Socioeconomic Status (0-10 books = 0)				
2. 11-25 books	0.94	1.34	-3.18*	1.61
3. 26-100 books	2.87*	1.36	-1.29	1.61
4. 101-200 books	3.99**	1.43	-2.64	1.76
5. More than 200 books	2.98	1.53	-2.16	1.77
6. Gender (girls = 0)	-1.15	0.80	1.91*	0.76

**Statistically significant for $p < 0.01$
*Statistically significant for $p < 0.05$

The results for science show that having "11-25 books" compared to "0-10 books" was associated with a significant decrease in the size of the mode effect ($\beta_2 = -3.18$, $p < 0.05$). After controlling for gender and digital self-efficacy, the size of the mode effect decreased for each for each category of socioeconomic status compared to the reference category. Relative to the international average mode effect of 8 score points on fourth grade science scores, this 3-point effect is relatively large.

The regression model results at the eighth grade, shown in Exhibit 4.16, were similar to the ANOVA results in Phase 1. However, similar to the fourth grade results, a significant effect was detected for socioeconomic status on science mode effects, where having "101-200 books" was negatively associated with the size of the score difference ($\beta_4 = -4.71$, $p < 0.01$), holding other predictor variables constant. Considering that the

international average score difference in science was 7 points, this effect was similarly large at the eighth grade.

Exhibit 4.16: International Average Regression Coefficients—Eighth Grade (11 countries)

	Mathematics		Science	
	B	***SE***	***B***	***SE***
Intercept	17.49**	2.75	13.42**	2.86
1. Digital Self-Efficacy	-0.30	0.24	-0.35	0.24
Socioeconomic Status (0-10 books = 0)				
2. 11-25 books	-2.07	1.67	-1.60	1.63
3. 26-100 books	-1.32	1.61	-2.40	1.57
4. 101-200 books	0.84	1.86	-4.71**	1.81
5. More than 200 books	1.67	1.99	0.16	2.00
6. Gender (girls = 0)	-1.11	1.49	-2.86	1.63

**Statistically significant for $p < 0.01$

The large effects found for socioeconomic status on science scores warranted further exploration of the relationships to ensure results are accurate. The ANOVA models did not detect a significant effect of socioeconomic status on science scores, so the significant effects detected by the regression model may only apply to girls with average levels of digital self-efficacy. To explore this further, a second regression model was run on the fourth grade and eighth grade science mode effects with only socioeconomic status included as a predictor variable.

Exhibit 4.17 shows the results for both models. The model intercepts at the fourth and eighth grades match the mean science difference scores for the students in the "0-10 books" categories presented in Exhibits 4.5 and 4.6, respectively. Additionally, the

decrease in the size of the mode effects compared to the reference category correspond to the each of the category differences in Exhibits 4.5 and 4.6. At the fourth grade, having "11-25 books" at home was associated with approximately a 3-point decrease in the difference between paperTIMSS and eTIMSS science scores, on average. At the eighth grade, being in this category was associated with almost a 5-point average decrease in the size of the science mode effect. These results suggest that the relationship between socioeconomic status and science mode effects is about the same across boys and girls with varying levels of digital self-efficacy.

Exhibit 4.17: International Average Regression Coefficients of Socioeconomic Status on Mode Effects in Science—Fourth and Eighth Grades

	Fourth Grade		**Eighth Grade**	
	B	***SE***	***B***	***SE***
Intercept (0-10 books)	10.17**	1.66	8.88**	1.71
1. 11-25 books	-3.24*	1.59	-1.62	1.65
2. 26-100 books	-1.53	1.65	-2.64	1.55
3. 101-200 books	-3.01	1.81	-4.91**	1.83
4. More than 200 books	-2.46	1.80	0.03	1.98

**Statistically significant for $p < 0.01$
*Statistically significant for $p < 0.05$

Discussion of the Results

The analysis of the eTIMSS Item Equivalence Database by subgroups sought to address whether the mode of administration differentially affected students based on gender, socioeconomic status, and level of digital self-efficacy. Descriptive analysis of paperTIMSS, eTIMSS, and difference scores by student subgroups revealed that, on average, the mode effect was mostly uniform across student subgroups by socioeconomic

status, gender, and digital self-efficacy. These results provide further evidence for the comparability of eTIMSS and paperTIMSS scores.

However, despite not being generalizable to the full TIMSS populations or population subgroups, the results also suggest that certain subgroups of students may be at risk for performing differently on eTIMSS versus paperTIMSS, particularly with regard to socioeconomic status. In Phase 1, the ANOVA models detected small effects of socioeconomic status on mathematics mode effects at the fourth and eighth grades (η_p^2 = 0.002). In addition, the Phase 2 results revealed a significant negative association between socioeconomic status and the size of science score differences, suggesting that higher socioeconomic status may has a positive impact on eTIMSS performance in science, regardless of students' gender and degree of digital self-efficacy.

However, results for socioeconomic status were conflicting across mathematics and science, where students with more books in their home exhibited larger performance differences in mathematics, but smaller performance differences in science, on average. Moreover, upon further analysis, no clear relationship emerged between socioeconomic status and mathematics mode effects. These results were inconsistent with prior research suggesting that students of low socioeconomic status may perform relatively worse on computer-based assessments (MacCann, 2006). The students in the low socioeconomic status categories did not have smaller differences between modes compared to the high categories.

A very small, but significant interaction between socioeconomic status and digital self-efficacy in predicting mathematics mode effects at the fourth grade (η_p^2 = 0.001) suggests that generally, students with lower levels of digital self-efficacy may be at risk

for exhibiting larger mode differences in achievement, particularly when these students are of lower socioeconomic status. This is consistent with prior research showing greater access to technology is associated with higher self-efficacy for using it (Lei & Zhou, 2012). The smallest score differences were seen for students with "High" digital self-efficacy. However, the effect of digital self-efficacy was non-significant in all four ANOVA models and all four regression models and overall the mode effects were approximately equal across groups within the margin of error. In addition, it is likely the relationships comprising the interaction effect with socioeconomic status in the ANOVA model are unreliable due to smaller samples in the low digital self-efficacy category, particularly when distributed across the five categories of socioeconomic status.

Phase 1 and 2 results showed similar results for the effect of gender. While results suggest that boys may be at more risk for performing worse on eTIMSS compared to paperTIMSS in science at the fourth grade, girls had larger average score differences than boys at the eighth grade in both subjects. In addition, mode effects varied only 2 points on average between boys and girls, even with digital self-efficacy and socioeconomic status held constant in the regression models. This is inconsistent with earlier studies finding that girls tend to perform worse on computer-based assessments (Gallagher et al., 2002; Jerrim, 206; Parshall & Kromrey, 1993). Overall relative achievement (regardless of mode) as well as sample size may have had an impact on the statistically significant results in both types of models.

Taken together, the predictor variables accounted for a small overall percentage of the variance between paperTIMSS and eTIMSS scores. Therefore, the analysis results suggest that mode of administration did not affect students differently according to these

characteristics on TIMSS, overall. However, students of low socioeconomic status may be at greater risk for exhibiting mode effects.

Limitations and Suggestions for Further Research

Although the within-subjects design of the eTIMSS Pilot / Item Equivalence Study made paperTIMSS and eTIMSS scores directly comparable, the results of analyses using the scores did not give results generalizable to the full TIMSS student populations or population subgroups. With small, non-random samples of students for each country, standard errors associated with the results were not accurate of population results. The jackknife standard errors only controlled for variance due to clustering of students in schools. Therefore, analysis on the country level was not feasible and valid conclusions can only be made about the effects of the background variables on the score distributions, as well as groups of students who may be at higher risk for exhibiting mode effects.

When analyzing large-scale assessment data, statistical significance is necessary but not sufficient evidence for a relationship in the data. Future analyses using nationally representative samples and using analysis techniques that further account for the distribution of variance within and across countries will help produce more reliable results. In particular, hierarchical linear models, or multi-level models, may better model the variance within and across countries in achievement, as well as the variance in the magnitude of the relationships between achievement and predictor variables. Furthermore, with nationally representative samples, a structural equation modelling approach could help further disentangle the relationships among socioeconomic status, gender, and digital self-efficacy in predicting mode effects and how these relationships vary across countries.

Chapter 5: Discussion

Overview of Dissertation

This dissertation describes the steps in developing a strategy for maintaining the TIMSS 20-year trend measurements in TIMSS 2019. About half of the 60 participating countries are transitioning to eTIMSS, a new computer- and tablet-based mode of assessment delivery.

Four major initiatives took place over a three-year period:

- Developing reliable and secure software and application components of the eAssessment system for TIMSS that accommodates diversity in the types of digital devices available within countries
- Designing a user interface for eTIMSS to facilitate the student assessment experience and prevent difficulties in navigating the assessment or inputting responses
- Converting 400 of the 2015 assessment items that will be brought forward, called trend items, to the eTIMSS interface, while working to maintain measurement equivalence as much as possible
- Conducting studies to examine the impact of the mode of administration on the measurement properties of the trend items

The development efforts were informed by the experiences of other large-scale assessment programs (e.g., NAEP, PISA) as well as prior research on mode effects and user interface design. AIR Cognitive Labs were conducted in August–September 2015, the eTIMSS prePilot in September 2016, and the eTIMSS Pilot / Item Equivalence Study

in May–June 2017. Further analysis of the eTIMSS Item Equivalence Database by student subgroups provided further information about the nature of the mode effects in relation to socioeconomic status, gender, and digital self-efficacy.

eTIMSS Pilot / Item Equivalence Study

The eTIMSS Pilot / Item Equivalence Study was administered in 25 countries beginning in May 2017. The study had two primary purposes:

- Tryout the eTIMSS systems for the first time on a large-scale to inform further development, including the eTIMSS Translation System, Player, Data Monitor, and Scoring System
- Use the trend items to examine the extent to which items behave the same in paper and electronic modes

The second purpose involved conducting an item equivalence study to examine the effect of the new mode of administration on the measurement properties of the trend items and scale scores to determine if enough of the TIMSS trend items were "equivalent" to be the basis for linking eTIMSS to paper-based trends. The same students experienced the trend items in both paperTIMSS and eTIMSS in a counterbalanced design that prevented students from receiving the same items in both modes. Half of students took paperTIMSS first, and half of students took eTIMSS first. Sample sizes for analysis included 16,894 fourth grade students and 9,164 eighth grade students.

Analysis of the eTIMSS Item Equivalence Database

The eTIMSS Item Equivalence Database includes data for 24 fourth grade countries and 13 eighth grade countries. Each student has four sets of five plausible values for achievement scores in mathematics and science for both paperTIMSS and

eTIMSS. The database includes paperTIMSS and eTIMSS data for 159 items at the fourth grade (77 mathematics items and 82 science items) and 207 items at the eighth grade (97 mathematics items and 110 science items).

Three phases of analysis were conducted for the eTIMSS Pilot / Item Equivalence Study. As described in Chapter 3, Phase 1 included an a priori analysis of item equivalence to classify the trend items according to their differences across paper and digital formats. In Phase 2, an item-by-item review analyzed the differences between paperTIMSS and eTIMSS item statistics by country and overall to examine measurement equivalence based on item difficulty (percent correct), item discrimination (point-biserial correlations), and percent missing ("omitted" and "not reached") statistics for each trend item. Average item statistics also were examined by digital item type. Lastly, Phase 3 examined the effect of the mode of administration on the TIMSS scale scores.

The eTIMSS Item Equivalence Database also includes data collected from the eTIMSS questionnaire which provides information about students' socioeconomic status, gender, and their experiences and attitudes with using computers and tablets. The review of the mode effect literature summarized in Chapter 4 suggests that mode effects can vary according to these student attributes. Further analysis of mean paperTIMSS and eTIMSS scores across student subgroups allowed for further examining score comparability across modes. An analysis of variance and a multiple linear regression analysis explored the degree to which students' socioeconomic status, gender, and digital self-efficacy might explain the differences between paperTIMSS and eTIMSS scores.

Major Findings

Taken together, the results of the eTIMSS Pilot / Item Equivalence Study and the analysis conducted for this dissertation led to five major conclusions.

1. **Converting the paper trend items to a digital format was mostly successful, with some improvements needed.**

Despite the awareness of TIMSS & PIRLS International Study Center staff that some of the trend items may cause frustration for students in the eTIMSS Pilot / Item Equivalence Study, the a priori item analysis indicated that efforts to keep the trend items looking the same across paper and digital formats were mostly successful. Overall, the results provided face validity evidence that the mathematics and science constructs assessed by the paper trend items were maintained in the majority of the trend items after the conversion to their digital formats.

The difference was more pronounced in response modes for constructed response items. These primarily included items for which students had to draw or write on screen and "number pad" items requiring the use of the on-screen number keypad for inputting numerical responses. Unfortunately, the number keypad was not completely functional for the eTIMSS Pilot / Item Equivalence Study.

Upon further review of item statistics, more items were identified as problematic in their digital formats, including several items with keyboard entry boxes in tables that were too small for character-based languages. A few items with severe scrolling requirements (usually comprising two pages on paper) had omit rates as high as 50 percent for eTIMSS, substantially higher than omit rates for paperTIMSS.

These issues resulted in items originally classified as *expected non-invariant* being deleted from the item equivalence analysis. This included 28 items at the fourth grade (15 mathematics items and 13 science items) and 25 items at the eighth grade (17 mathematics items and 8 science items). Nevertheless, the remaining sample sizes for the analysis were sufficient, with 77 mathematics items and 82 science items at the fourth grade, and 97 mathematics items and 110 science items at the eighth grade.

2. There is evidence for a general mode effect on the difficulty of the trend items, especially in mathematics.

Exhibit 5.1 presents the summary of the results from the eTIMSS Pilot / Item Equivalence Study, which shows average mode effects for the fourth and eighth grades in mathematics and science, respectively. International average percent correct values reflect the average difference in percent correct statistics (item difficulty) between paperTIMSS and eTIMSS. The achievement score means reflect the average difference between paperTIMSS and eTIMSS in scale score points.

The results provide evidence for a general mode effect impacting all trend items, on average. The effect on the mathematics scale scores was particularly substantial for the TIMSS context. These findings led to the conclusion that the trend items could not form the basis for the link between paperTIMSS and eTIMSS to maintain trends. The mode effect favored paper items, with higher percent correct values on paperTIMSS compared to eTIMSS, on average. Some items exhibited little to no mode effects, but some had much larger mode effects. Nevertheless, there were not enough trend items equivalent across modes to rely on common item equating for linking paperTIMSS and eTIMSS to maintain trends.

Exhibit 5.1: Summary of the Results of the eTIMSS Pilot / Item Equivalence Study

	International Average Mode Effects (paperTIMSS – eTIMSS Differences)	
	Percent Correct (*SD*)	**Achievement Score (*SE*)**
Fourth Grade (21 countries)		
Mathematics	3.6 (6.0)	14 (0.7)
Science	1.7 (6.0)	8 (0.6)
Eighth Grade (11 countries)		
Mathematics	3.4 (5.3)	14 (1.0)
Science	1.5 (5.7)	7 (1.0)

( ) Standard deviations for percent correct and standard errors for achievement scores appear in parentheses. Because of rounding some results may appear inconsistent.

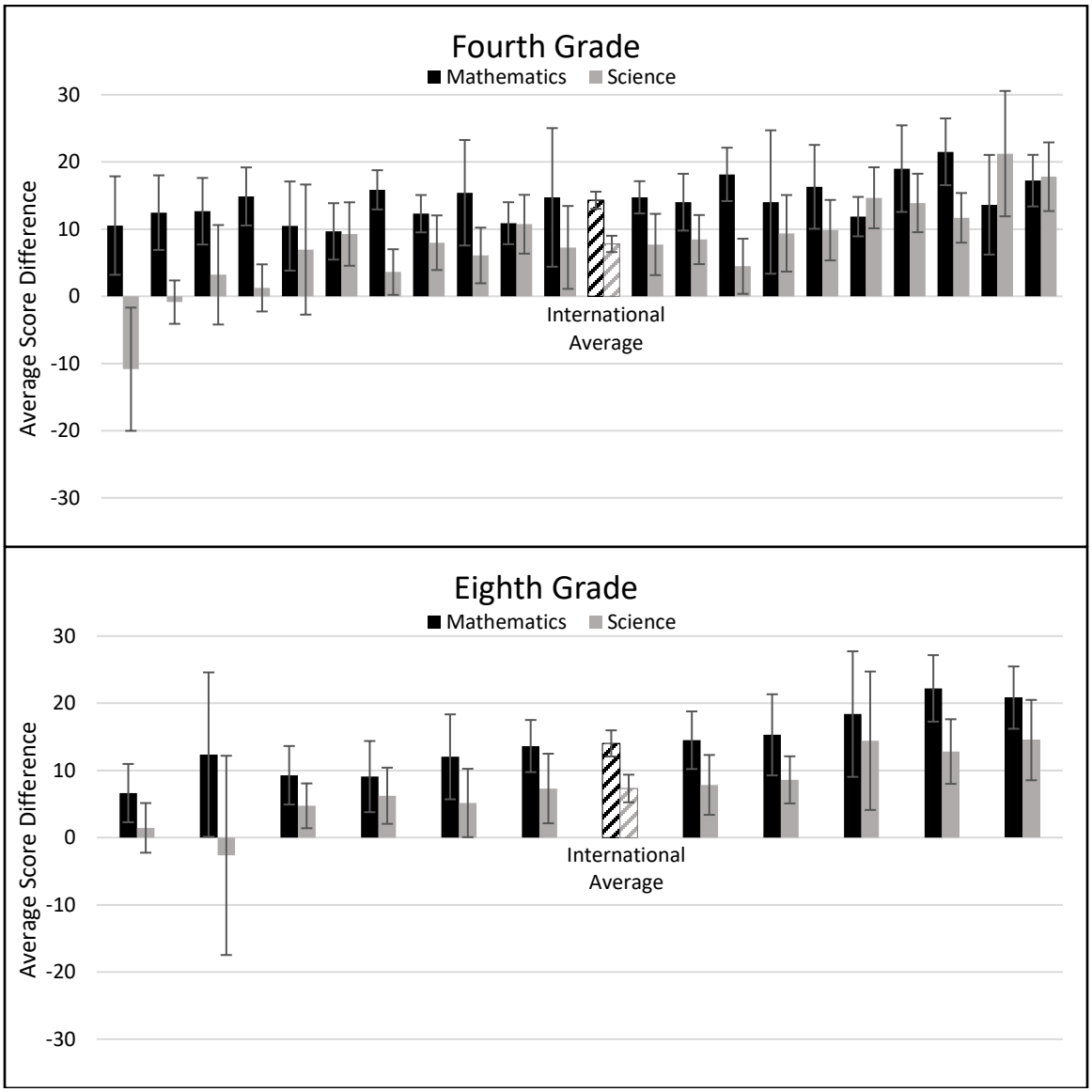
The relatively larger average item mode effects in mathematics (3.6% at the fourth grade and 3.4% at the eighth grade) compared to science (1.7% and 1.5%, respectively) could be further understood in the scale score analysis with scores in the TIMSS metric. International average differences of 14 scale score points between paperTIMSS and eTIMSS were found in mathematics at the fourth grade and eighth grade. In science, an 8-point international average difference was found at the fourth grade and a 7-point difference was found at the eighth grade.

The differences between the mode effects in mathematics and science are hypothesized to be related to the nature of the items in the respective pools. Science trend items at each grade consisted primarily of multiple-choice and keyboard item types, whereas mathematics had mostly number pad items. About one-third of the mathematics items required number pad inputs: 25 out of 77 items at the fourth grade and 29 out of 97 items at the eighth grade.

3. There was some variation in the mode effects across countries.

Exhibit 5.2 presents bar charts illustrating the size of the average mathematics and science score differences by country for the fourth and eighth grades, respectively. For each grade, the countries appear in order by their average mode effect size across mathematics and science. Negative values indicate that performance on eTIMSS was better than performance on paperTIMSS, on average.

Exhibit 5.2: Country Distribution of Mathematics and Science Score Mode Effects



Errors bars represent 95% confidence intervals for estimated country difference scores.

Although interpretation of country-level results is limited without nationally representative samples, the size of the 95 percent confidence intervals suggest that much of the variance across countries may be negligible after accounting for sampling error. However, the results of the scale score analysis conducted for the eTIMSS Pilot / Item Equivalence Study indicate there is some variation in the mathematics and science mode effects across countries.

4. **Overall, the mode effects were not related to student characteristics theorized to influence computer-based test performance.**

The results of the analysis of the eTIMSS Item Equivalence Database in Chapter 4 suggest that, overall, the mode effects on the trend items affected students uniformly across subgroups of students based on socioeconomic status, gender, and digital self-efficacy. These student characteristics explained a negligible proportion of the variance in achievement score differences between paperTIMSS and eTIMSS. However, the results suggest that some groups of students should be considered more "at risk" for performing differently on eTIMSS compared to paperTIMSS, particularly students of lower socioeconomic status.

5. **The mathematics and science constructs measured by the trend items are equivalent for paperTIMSS and eTIMSS.**

With analysis by student subgroups showing that the mode of administration affected students uniformly, the results in this dissertation suggest that, despite the clear presence of the mode effects, the TIMSS mathematics and science constructs are essentially the same whether assessed by paperTIMSS or eTIMSS. Therefore, paperTIMSS and eTIMSS scores can be made comparable, and the differences between

the scores that resulted from the mode effects on the trend item difficulties can be corrected through equating (Winter, 2010).

The results of the a priori item analysis conducted for the eTIMSS Pilot / Item Equivalence Study provided face validity evidence for the equivalence of the mathematics and science constructs measured by the trend items. The quantitative item analysis results also showed negligible differences in item discrimination statistics between paperTIMSS and eTIMSS, with less than 0.03 average differences in point-biserial correlation coefficients for each subject and grade. At the score level, cross-mode correlation coefficients reflecting the relationships between paperTIMSS and eTIMSS scores were large ($r > 0.95$). Lastly, examining mean scores by country for each grade and subject indicated that country rankings were about the same for paperTIMSS and eTIMSS—with no differences in rankings at the high and low ends of the score distribution.

Preserving TIMSS Trend Measurements: Next Steps

Linking paperTIMSS and eTIMSS with Common Population Equating

The differences in the difficulty of the trend items between paperTIMSS and eTIMSS require a modified approach for maintaining trend measurements in TIMSS 2019 compared to the one used in previous cycles. Typically, trend items administered across multiple assessments allow for TIMSS IRT scales to be linked from cycle to cycle through a concurrent calibration process that places the item parameters from the newly collected data on the same scale as the item parameters from the previous cycle. Then,

linear transformations are applied to the student proficiency scores to place them on the same scale as the previous assessment.

The differences in trend item parameters between subsequent assessments are usually very small, and change only slightly due to the presence of new items and new countries in the trend pool. However, the results of the eTIMSS Pilot / Item Equivalence Study showed that the differences in the difficulty of trend items between paperTIMSS and eTIMSS are too large to rely on the common item equating approach, especially in mathematics. Instead, a common population equating approach will be used to place paperTIMSS and eTIMSS on the same scale for TIMSS 2019.

Bridge Samples

With the plan for collecting bridge data in 2019, the common population approach and a two-stage linear transformation (illustrated in Exhibit 3.26) will correct for the differences in trend item parameters compared to past cycles as well as the differences in the proficiency distributions between paperTIMSS and eTIMSS that occur due to mode effects on the trend items. This equating approach proved successful when a new booklet design was introduced for TIMSS 2007 (Foy, Galia, & Li, 2008).

As part of TIMSS 2019, each eTIMSS trend country will administer the trend items in paper booklets to a nationally representative "bridge sample" of students—an additional 25 percent of students in addition to the full eTIMSS sample. Then, the bridge data will be calibrated concurrently with the pool of data collected from paperTIMSS countries and the paper trend data from TIMSS 2015. A linear transformation will place all paperTIMSS 2019 results (including bridge data) on the same scale as TIMSS 2015.

Data for eTIMSS will undergo a separate calibration, and the second linear transformation will place eTIMSS on the same scale as paperTIMSS.

Improving the Trend Items

The mode effects may also decrease after improvements are made to the trend items. Many of these efforts have already begun and a better version of the number pad has been included in the TIMSS 2019 Field Test. Development of a line tool also is underway for use in items that require students to draw graphs and diagrams. Efforts to keep item images a reasonable size to prevent scrolling have been successful thus far, and research is underway to identify additional solutions.

Re-assessing Trend Item Equivalence

The size of the mode effect also may be reduced after improvements to the eAssessment system and with nationally representative samples of students. Therefore, the trend items will be re-assessed for mode effects following data collection in 2019 to re-estimate the mode effect. The use of bridge samples will allow for an updated analysis of item equivalence by the TIMSS & PIRLS International Study Center. Countries who participate in eTIMSS 2019 may also conduct their own item equivalence analyses with their bridge data to examine the particular nuances of the mode effect for students in their country.

Additional Recommendations for TIMSS

Many improvements for eTIMSS are already in place for the TIMSS 2019 Field Test. The results of the TIMSS 2019 Field Test should be used to help further improve the experience for students to reduce the potential for construct irrelevant variance in

achievement scores. This research identifies three ways to help mitigate the impact of the students' familiarity with computers or tablets on their eTIMSS performance—1) improve the assessment directions; 2) collect data about students' user experience with eTIMSS; and 3) continue to measure students' computer and tablet familiarity and attitudes.

eTIMSS Direction Module

This dissertation research highlighted the importance of students being able to implement the types of response actions required to take the eTIMSS assessment. Several other researchers have similarly suggested that familiarizing students with using computers and tablets in the classroom may help mitigate performance differences on paper versus digital assessments (Davis et al., 2017; Strain-Seymour et al., 2013). These findings led to the addition of a 10-screen directions module for the TIMSS 2019 Field Test, where students practice navigating through the assessment and answering the various item types. This opportunity for practice should help reduce some difficulties experienced by students and result in better measurement of the mathematics and science constructs.

Collecting User Experience Data

The eTIMSS questionnaire administered to students for the Field Test asks students about their eTIMSS experience. Students are asked whether they liked that the assessment was given on a computer or tablet, as well as if they experienced any technical difficulties while taking the assessment, including software issues, or difficulty navigating the user interface or responding to items. TIMSS should use this data to

improve the assessment for students, with the goal that students attend to test content rather than become pre-occupied with navigating the assessment.

Measuring Computer Experience and Attitudes

Because of the possible detriment of low levels of digital self-efficacy on students taking eTIMSS, TIMSS should measure the digital self-efficacy construct and other measures of computer and tablet competence. After preliminary analysis of the item statistics from the eTIMSS Pilot / Item Equivalence Study, the digital self-efficacy scale was revised for inclusion in the eTIMSS questionnaire for the TIMSS 2019 Field Test.

Because research suggests that the positive effects of technology experience on test performance may be more highly related to experience using technology for educational reasons in school settings than, for example, playing computer games (Kubiatko & Vlkova, 2010), scales about students' technology familiarity and self-efficacy should ask about uses or particular actions of digital devices that are relevant to how computers and tablets are used in educational or assessment contexts.

With the goal of constructing a scale with a positive relationship with achievement, items measuring these constructs should not apply to remedial situations in school, because students who report engaging in such activities could be more likely to have lower achievement. While such questions are important to ask, the data are unlikely to contribute to a construct predictive of computer-based assessment performance for students of all abilities.

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER ConQuest: Generalised item response modelling software* [Computer software]. Version 4. Camberwell, Victoria: Australian Council for Educational Research.

American Institutes for Research (AIR). (2013, September). *Smarter Balanced Assessment Consortium: Cognitive laboratories technical report*. Prepared for the Smarter Balanced Assessment Consortium. Washington, DC: Author.

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.

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessment (APA). (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC: Author.

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

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education, 16*(3), 191-205.

Buerger, S., Kroehne, U., & Goldhammer, F. (2016). The transition to computer-based testing in large-scale assessments: Investigating (partial) measurement invariance between modes. *Psychological Test and Assessment Modeling, 58*(4), 597-616.

Bundsgaard, J., & Gerrick, J. (2017). Patterns of students' computer use and relations to their computer and information literacy: results of a latent class analysis and implications for teaching and learning. *Large-scale Assessments in Education, 5*(16), 1-15.

Chen, G., Cheng, W., Chang, T-W, Zheng, X., & Huang, R. (2014). A comparison of reading comprehension across paper, computer screens, and tablets: Does tablet familiarity matter? *Journal of Computers in Education, 1*(3), 213-225.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York, NY: Routledge Academic.

Cooper, J. (2006). The digital divide: The special case of gender. *Journal of Computer Assisted Learning, 22*, 320-334.

Davis, L. L., Kong, X., McBride, Y., & Morrison, K. (2017). Device comparability of tablets and computers for assessment purposes. *Applied Measurement in Education, 30*(1), 16-26.

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

Duque, M. (2016). *Is there a PARCC mode effect?* Strategic Data Project Fellowship Capstone Report. Center for Education Policy Research, Harvard University. Retrieved from https://sdp.cepr.harvard.edu/

Falloon, G. (2013). Young students using iPads: App design and content influences on their learning pathways. *Computers & Education, 68*, 505-521.

Foy, P., Galia, J., & Li, I. (2008). Scaling the data from the TIMSS 2007 mathematics and science assessments. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Foy, P., & LaRoche, S. (2016). Estimating standard errors in the TIMSS 2015 results. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and proecudres in TIMSS 2015* (pp. 4.1-4.69). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timss.bc.edu/publications/timss/2015-methods/chapter-4.html

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

Foy, P., & Yin, L. (2016). Scaling the TIMSS 2015 achievement data. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 13.1-13.62). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timss.bc.edu/publications/timss/2015-methods/chapter-13.html

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. *Journal of Educational Measurement, 39*(2), 133-147.

Grzanna, M. (2017, December 4). Computer instead of paper. *Süddeutsche Zeitung*. Retrieved from http://www.sueddeutsche.de/bildung/pisa-computer-statt-papier-1.3277018

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

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy, & Practice, 23*(4), 495-518.

Jerrim, J. (2018). *A digital divide? Randomized evidence on the impact of computer-based assessment in PISA*. CfEE Research Brief. Retrieved from http://www.cfee.org.uk/sites/default/files/CfEE%20Digital%20Divide_1.pdf

Johansone, I. (2016). Survey operations procedures in TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 6.1-6.22). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timss.bc.edu/publications/timss/2015-methods/chapter-6.html

Johnson, M., & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, and Assessment, 4*(5), 1-35.

Kingston, N. M. (2008). Comparability of computer- and paper- administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22*(1), 22-37.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746-759.

Kubiatko, M., & Vlckova, K. (2010). The relationship between ICT use and science knowledge for Czech students: A secondary analysis of PISA 2006. *International Journal of Science and Mathematics Education, 8,* 523–543.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 1-12.

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

Lei, J., & Zhou, J. (2012). Digital divide: How do home internet access and parental support affect student outcomes? *Education (Basel), 2*, 45-53.

Linacre, J. M. (2017). *WINSTEPS Rasch measurement* [computer program]. Beaverton, Oregon.

Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Luu, K., & Freeman, J.G. (2011). An analysis of the relationship between information and communication technology (ICT) and scientific literacy in Canada and Australia. *Computers & Education, 56,* 1072–1082.

MacCann, R. (2006). The equivalence of online and traditional testing for different subpopulations and item types. *British Journal of Educational Technology, 37*(1), 79-81.

Martin, M. O., Mullis, I. V. S., Beaton, A. E., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1998). *Science achievement in the primary school years: IEA's third international mathematics and science report.* Chestnut Hill, MA: Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O. Mullis, I. V. S., & Foy, P. (2013). TIMSS 2015 assessment design. In I. V. S. Mullis & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 85-98). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Foy, P. & Hooper, M. (Eds.). (2016a). TIMSS achievement methodology. In *Methods and procedures in TIMSS 2015* (pp. 12.1-12.9). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timss.bc.edu/publications/timss/2015-methods/chapter-12.html

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016b). *TIMSS 2015 international results in mathematics*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., Fishbein, B., & Liu, J. (2017). Creating and interpreting the PIRLS 2016 context questionnaire scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 14.1-14.106). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-14.html

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 427-450.

Mayer, R. E. (2009). *Multimedia learning* (2nd Ed.). New York: Cambridge University Press.

Mayer, R. E. (2014). Research-based principles for designing multimedia instruction. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying science of learning into education: Infusing psychological science into the curriculum.* Retrieved from https://pdfs.semanticscholar.org/e21c/e17f7f04962986c5e2733f18662a4da47c9b.pdf

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature.* College Board Rep. No. 88-8, ETS RR No. 88-21. Princeton, NJ: Educational Testing Service.

Mazzeo, J., & von Davier, M. (2008). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. *Education Working Papers EDU/PISA/GB (2008), 28*, 23-24.

Mazzeo, J., & von Davier, M. (2014). Linking scales 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. 229-258). Boca Raton, FL: Chapman & Hall, CRC Press.

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education, 39*, 299-312.

Meyer, S., Cockle, M., & Taneva, M. (2016). Creating the TIMSS 2015 International Database. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 10.1-10.12). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timss.bc.edu/publications/timss/2015-methods/chapter-10.html

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177-196.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2013). *TIMSS 2015 assessment frameworks.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

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., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1998). *Mathematics achievement in the primary school years: IEA's third international mathematics and science report.* Chestnut Hill, MA: Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/

Mullis, I. V. S., Martin, M. O., & Hooper, M. (2017). Measuring changing educational contexts in a changing world: Evolution of the TIMSS and PIRLS questionnaires. In M. Rosén, K. Y. Hansen, & U. Wolff (Eds.), *Cognitive abilities and educational outcomes* (pp. 207-222). Methodology of Educational Measurement and Assessment. Switzerland: Springer International Publishing.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176.

Muraki, E., & Bock, R. D. (1991). *PARSCALE* [computer software]. Lincolnwood, IL: Scientific Software International.

O'Dwyer, L. M., Russell, M., Bebell, D., & Tucker-Seeley, K. R. (2005). Examining the relationship between home and school computer use and students' English/language arts test scores. *Journal of Technology, Learning, and Assessment, 3*(3), 1-45.

OECD. (2015). *Students, computers, and learning: Making the connection*. PISA, OECD Publishing. Retrieved from http://www.oecd.org/about/publishing/corrigenda.htm

OECD. (2016). Annex A5: Changes in the administration and scaling of PISA 2015 and implications for trends analyses. In *PISA 2015 results (Volume I): Excellence and equity in education* (pp. 305-317). Paris: PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264266490-en

OECD. (2017). Computer platform / Computer-based tests. *PISA 2015 technical report*. Retrieved from http://www.oecd.org/pisa/data/2015-technical-report/

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*, 315-333.

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing, 14*, 1-21.

PARCC. (2017, January). *Technology guidelines for PARCC assessments v6.1*. Retrieved from http://www.parcconline.org/technology

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

Parshall, C. G., & Kromrey, J. D. (1993, April). *Computer testing versus paper-and-pencil testing: An analysis of examinee characteristics associated with mode effect*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer-Verlag.

Pearson, E. S., & Hartley, H. O. (1954) *Biometrika tables for statisticians, Volume I.* New York: Cambridge University Press.

Pisacreta, D. (2013, June). *Comparison of a test delivered using an iPad versus a laptop computer: Usability study results.* Paper presented at the Council of Chief State School Officers (CCSSO) National Conference on Student Assessment (NCSA), National Harbor, MD.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. (2005). A comparative evaluation of score results form computerized and paper & pencil mathematics testing in a alrge scale state assessment program. *Journal of Technology, Learning, and Assessment, 4*(6). Retrieved from http://www.jtla.org

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passaged-based tests. *Journal of Technology, Learning, and Assessment, 2*(6), 1-45.

Pruet, P., Ang, C.S., & Farzin, D. (2016). Understanding tablet computer usage among primary school students in underdeveloped areas: Students' technology experience, learning styles and attitudes. *Computers in Human Behavior, 55,* 1131–1144.

Randall, J., Sireci, S., Li, X., & Kaira, L. (2012). Evaluating the comparability of paper- and computer-based science tests across sex and SES subgroups. *Educational Measurement: Issues and Practice, 31*(4), 2-12.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics, 4*(3), 207-230.

Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F., & Heine, J-H. (2016). Challenges in the estimation of trends in school performance studies: A scaling of the German PISA data. *Diagnostica.* DOI: *http://dx.doi.org/10.1026/0012-1924/a000177*

Rogers, A., Tang, C., Lin, J.-J., & Kandathil, M. (2006). *DGROUP* [computer software]. Princeton, NJ: Educational Testing Service.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(2). Retrieved from http://epaa.asu.edu/epaa/v7n20/

Russell, M. (2002). *The influence of computer-print on rater scores.* Chestnut Hill, MA: Technology and Assessment Study Collaborative, Boston College.

Russell, M., Goldberg, A., & O'Connor, K. (2003). *Computer-based testing and validity: A look back and into the future.* Chestnut Hill, MA: Technology and Assessment Study Collaborative, Boston College.

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.

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, D.C.: U.S. Government Printing Office.

Spiezia, V. (2010). *Does computer use increase educational achievements? Student-level evidence from PISA*. OECD Journal: Economic Studies.

Skryabin, M., Zhang, J., Liu, L., & Zhang, D. (2015). How the ICT development level and usage influence student achievement in reading, mathematics, and science. *Computers & Education, 85*, 49-58.

Smarter Balanced Assessment Consortium (SBAC). (2014, August). *The Smarter Balanced technology strategy framework and testing device requirements.* Prepared by Navigation North Learning. Retrieved from http://smarterbalanced.org

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Strain-Seymour, E., Craft, J., Davis, L. L., & Elbom, J. (2013, July). *Testing on tablets: Part I of a series of usability studies on the use of tablets for K-12 assessment programs.* Pearson White Paper. Retrieved from http://researchnetwork.pearson.com/

Thissen, D., & Norton, S. (2013). *What might changes in psychometric approaches to statewide testing mean for NAEP?* Commissioned by the NAEP Validity Studies Panel. Retrieved from http://files.eric.ed.gov/fulltext/ED545241.pdf

Tømte, C., & Hatlevik, O.E. (2011). Gender-differences in self-efficacy ICT related to various ICT user profiles in Finland and Norway. How do self-efficacy, gender and ICT-user profiles relate to findings from PISA 2006. *Computers and Education, 57,* 1416–1426.

Wang, S., Jiao, H., Young, M. J., Brooks, T., Olson, J. (2007). *Educational and Psychological Measurement, 67*(2), 219-238.

Ward, H. (2017, March 24). Exclusive: PISA data may be incomparable, Schleicher admits. *TES*. Retrieved from https://www.tes.com/news/school-news/breaking-news/exclusive-pisa-data-may-be-incomparable-schleicher-admits

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450.

Way, D. W., Davis, L. L., Keng, L., & Strain-Seymour, E. (2016). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 260-284). New York and London: Taylor & Francis, Routledge.

Winter, P. C. (Ed.). (2010). *Evaluating the comparability of scores from achievement test variations.* Washington, DC: Council of Chief State School Officers.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Zhang, T., Xie, Q., Park, B. J., Kim, Y. Y., Broer, M., & Bohrnstedt, G. (2016). *Computer familiarity and its relationship to performance in three NAEP digital-based assessments* (AIR-NAEP Working Paper #01-2016). Washington, DC: American Institutes for Research.

Zhong, Z-J. (2011). From access to usage: The divide of self-reported digital skills. *Computers & Education, 56*, 736-746.

Appendix A: Constructing the Digital Self-Efficacy Scale

Selecting Scale Items

Before item calibration, the data were recoded for analysis and summary statistics were produced to make final selections of the scale items. The response categories for the items were recoded so that 3 = "Agree a lot" or "I definitely can;" 2 = "Agree a little" or "I probably can;" 1 = "Disagree a little" or "I probably cannot;" and 0 = "Disagree a lot" or "I definitely cannot." This allowed for interpreting higher scale scores as higher levels of the construct. Then, classical item statistics were produced using SPSS to examine the response patterns for each of the items. The statistics were weighted with SENWGT so each country contributed equally to the results. Cases were excluded for students who did not have responses to at least two of the items.

Exhibit A.1 presents the weighted international summary statistics for the eight items measuring digital self-efficacy based on responses from fourth grade students. The item numbers in the far left column correspond to those in Exhibit 4.4. The table is similar to the item almanacs produced for achievement items in Chapter 3, with the number of valid cases, item difficulty (DIFF), item discrimination (DISC), and percentages responding in each category (P_0, P_1, P_2, P_3, and P_M for missing). In addition, the table includes point-biserial correlations for each category (PB_0, PB_1, PB_2, PB_3, and PB_M for missing).

Exhibit A.1: International Summary Statistics for Items Measuring Digital Self-Efficacy—Fourth Grade

Item	Valid Cases	DIFF	DISC	Percentages					Point-Biserial Correlations				
				P_0	P_1	P_2	P_3	P_M	PB_0	PB_1	PB_2	PB_3	PB_M
1A	9,988	82.5	0.63	2.8	6.4	31.1	59.6	2.6	-0.35	-0.34	-0.26	0.54	-0.03
1B	9,853	78.0	0.65	4.1	10.1	33.6	52.2	3.9	-0.37	-0.38	-0.18	0.55	-0.02
1C	9,878	80.4	0.66	3.9	8.9	29.4	57.8	3.7	-0.38	-0.38	-0.22	0.57	-0.02
2A	9,849	83.8	0.60	3.5	5.9	26.3	64.3	3.9	-0.34	-0.33	-0.27	0.54	-0.02
2B	9,769	87.1	0.57	3.3	5.3	18.4	73.1	4.7	-0.32	-0.31	-0.28	0.53	-0.02
2C	9,681	74.6	0.62	6.1	14.0	29.9	50.0	5.6	-0.37	-0.34	-0.13	0.54	-0.02
2D	9,646	68.5	0.58	9.6	16.2	33.3	40.9	5.9	-0.40	-0.28	-0.02	0.47	-0.03
2E	9,731	86.9	0.65	3.1	5.3	19.4	72.2	5.1	-0.38	-0.34	-0.33	0.60	-0.03

The results show that most of the items had very small percentages of students answering in the lowest category (P_0). This may be because of the purposive sample of students drawn for the study. One item (1B) had reversal in answer category behavior, with higher point-biserial correlations for the lowest answer category (0) compared to category 1. However, the difference in correlation between the two categories was small. All items had overall point-biserial correlations above 0.56, indicating that the items discriminate well among levels of the construct.

Exhibit A.2 presents the international summary statistics for the digital self-efficacy items at the eighth grade. Similar to the fourth grade, most of the items had very small percentages of students in the two low categories. Items 1A, 1C, 2A, 2B, and 2D had reversal in answer category behavior, with higher point-biserial correlations for the lower answer categories compared to category 2 (“Agree a little” and “I probably can”). All items had overall point-biserial correlations above 0.60.

Exhibit A.2: International Summary Statistics for Items Measuring Digital Self-Efficacy—Eighth Grade

Item	Valid Cases	DIFF	DISC	Percentages					Point-Biserial Correlations				
				P_0	P_1	P_2	P_3	P_M	PB_0	PB_1	PB_2	PB_3	PB_M
1A	5,313	76.8	0.69	2.8	10.5	40.3	46.4	1.1	-0.38	-0.40	-0.20	0.57	0.00
1B	5,302	74.5	0.72	3.9	13.3	37.9	44.8	1.3	-0.42	-0.40	-0.17	0.60	0.01
1C	5,304	85.1	0.65	1.5	5.1	30.0	63.4	1.3	-0.33	-0.34	-0.34	0.57	0.00
2A	5,287	88.8	0.68	1.3	3.6	22.6	72.5	1.6	-0.34	-0.37	-0.39	0.61	-0.03
2B	5,271	92.6	0.61	1.2	2.5	13.8	82.5	1.9	-0.31	-0.30	-0.39	0.57	-0.02
2C	5,254	73.0	0.63	5.8	16.0	31.9	46.4	2.2	-0.37	-0.34	-0.13	0.55	-0.02
2D	5,238	67.6	0.62	7.2	19.9	35.7	37.1	2.5	-0.39	-0.32	-0.03	0.50	-0.04
2E	5,261	92.8	0.62	1.2	1.9	14.2	82.7	2.1	-0.33	-0.26	-0.43	0.59	-0.02

Careful examination and comparison of the results at the fourth and eighth grades suggested that two items did not fit well with the others—item 2C (“Type using the correct fingers”) and item 2D (“Draw a picture using a computer”). These items did not fit well substantively with the others in terms of the target construct. Most students purposively sampled for the study were already familiar with typing, and item 2A also asked about “writing” on a computer. In addition, Most “canvas” item types requiring students to draw were deleted from the database, making the “draw a picture” item not as relevant in the assessment context. Students may also have interpreted the item to refer to creating a graphic, rather than using a mouse, finger, or stylus to draw on the screen. These items appeared to behave differently from the other items, with relatively higher omit rates at both grades and low discrimination statistics, particularly at the fourth grade. Further exploratory analysis revealed that removing the two items increased the percent

of variance accounted for by the items and resulted in scales more strongly related with achievement. Therefore, the final scales only included the six other items.

Evaluating Unidimensionality

The scales were validated using SPSS to confirm that the scale items constitute a single, underlying latent construct of students' digital self-efficacy. According to guidelines for TIMSS and PIRLS scale validation and statistical assumptions of Rasch models (Masters, 1982), background scales should be unidimensional. Unidimensionality of all six scale items was assessed in SPSS through a principal components analysis with pairwise deletion, based on only the cases included in IRT scaling. Reliability of the scale was also assessed by computing Cronbach's alpha reliability coefficient.

According to Reckase (1979), a unidimensional scale has a single dominant factor that accounts for approximately 20 percent of the variance, with no second dominant factor. Most TIMSS and PIRLS scales have a single dominant factor accounting for approximately 50 percent of the variance or higher (Martin, Mullis, Hooper, Yin, Foy, & Fishbein, 2017; Martin, Mullis, Hooper, Yin, Foy, & Palazzo, 2016). To evaluate unidimensionality, the principal components analysis first was restricted to a single component to evaluate whether the scale is unidimensional. Then, a second component solution was estimated allowing extraction of all eigenvalues greater than 1 to examine the remaining variation among the items. The resulting component loadings were examined to determine whether all six items strongly contribute to the scale. The component loadings represent correlations of the component variables with the

underlying factor. Commonly accepted criteria for a strongly contributing item has a component loading of at least 0.30.

The results showed that the six items form a unidimensional and reliable scale at the fourth grade and eighth grade. Exhibit A.3 shows the component solution for the scales based on the fourth grade and eighth grade data. At the fourth grade, restricting the extraction to a single component resulted in a factor accounting for 44 percent of the variance. Allowing multiple components to be extracted again resulted in one factor accounting for 44 percent of the total variance among the items. Although this is below the ideal 50 percent of variance, the six items resulted in a Cronbach's alpha of 0.74, indicating strong reliability, and the items had high component loadings above 0.58 indicating they all contributed to the scale.

Exhibit A.3: Principal Components Analysis of the Digital Self-Efficacy Scale—Fourth and Eighth Grades

Item	Component Loadings	
	Fourth Grade	**Eighth Grade**
1A	0.666	0.701
1B	0.670	0.711
1C	0.721	0.714
2A	0.614	0.724
2B	0.586	0.681
2D	0.694	0.701

There were similar results at the eighth grade. Restricting the extraction to a single component resulted in a factor accounting for 50 percent of the variance. All variables had factor loadings greater than 0.68 indicating all items contribute to the scale. Allowing multiple components to be extracted resulted in two factors accounting for 67

percent of the total variance. However, the second component only accounted for 17 percent of the variance, providing support for a single dominant factor (Reckase, 1979). The eighth grade scale also showed to have strong reliability (Cronbach's $\alpha = 0.79$).

Calibrating the Items

The items were calibrated at each grade using Conquest based on the combined data from all countries to produce international item parameters, with each country contributing equally to calibration using the SENWGT. The data were fit to a Rasch partial credit model (Masters, 1982), which models the probability that a student will respond a certain way to an item (i.e., in a particular response category) based on their level of the digital self-efficacy construct. For example, students who tend to respond "Agree a lot" or "I definitely can" to the items in Exhibit 4.4 have a higher level of digital self-efficacy than students who tend to respond "Disagree a lot" or "I definitely can't" to the items. Equation (A.1) below models the probability that person n with location θ_n on the latent digital self-efficacy construct would respond to item i in response category x_i out of m_i possible categories:

$$P_{x_i}(\theta_n) = \frac{\exp\left(\sum_{j=0}^{x_i} \theta_n - (\delta_i + \tau_{ij})\right)}{\sum_{h=0}^{m_i} \exp\left(\sum_{j=0}^{h} \theta_n - (\delta_i + \tau_{ij})\right)}, \quad \text{(A.1)}$$

where $x_i = 0$, 1, 2, or $m_i = 3$ for the digital self-efficacy scale; δ_i is the location of item i on the latent construct; and τ_{ij} is the step parameter, or item threshold parameter, between the

respective pairs of subsequent response categories, which represents the relative difficulty of endorsing the subsequently higher category.

Exhibit A.4 shows the international item parameters for the digital self-efficacy scale at the fourth grade and Exhibit A.5 shows the item parameters for the eighth grade scale. For each item, the exhibits present the number of cases included; the delta parameter, or item location on the theta scale; the tau parameters which indicate the location of the step parameter, or transition location parameter, expressed in deviations from delta; and the Rasch infit item statistic indicating how well the data fit the model. Infit values above 1.3 indicate unexpected response patterns. The results show that the data fit the model well for all six items at both grades.

Exhibit A.4: International Item Parameters for the Digital Self-Efficacy Scale—Fourth Grade

Item	Cases	delta	tau_1	tau_2	tau_3	Infit
1A	16,057	-0.033	-0.519	-0.594	1.113	0.95
1B	15,840	0.271	-0.761	-0.401	1.163	1.00
1C	15,882	0.141	-0.617	-0.340	0.957	0.93
2A	15,863	-0.029	-0.244	-0.582	0.826	1.09
2B	15,741	-0.182	-0.097	-0.267	0.364	1.16
2E	15,678	-0.168	-0.135	-0.323	0.458	0.88

Exhibit A.5: International Item Parameters for the Digital Self-Efficacy Scale—Eighth Grade

Item	Cases	delta	tau_1	tau_2	tau_3	Infit
1A	8,868	0.645	-1.584	-0.513	2.097	0.98
1B	8,850	0.887	-1.564	-0.357	1.921	1.00
1C	8,856	-0.029	-1.065	-0.688	1.753	0.99
2A	8,826	-0.337	-0.746	-0.644	1.390	1.01
2B	8,801	-0.585	-0.387	-0.427	0.814	1.12
2E	8,785	-0.582	-0.047	-0.784	0.831	0.86

Assessing Item Dependence

Because of the repetitive nature of the some of the scale items and the relatively lower percentage of variance accounted for by the items at the fourth grade (44% vs. 50% at the eighth grade), Winsteps software was used to check for local dependence between items. Pearson product-moment correlations of person score residuals between each item were computed in the form of Yen's (1984; 1993) Q_3 statistic. This statistic is the correlation between performance on two items after accounting for each's students overall performance on the scale.

Q_3 statistics are usually negative when data are unidimensional because an item score is included in both of the terms used to calculate the residual, and positive correlations may similarly underestimate the strength of the relationship. Therefore, Yen recommends making a small positive adjustment to the average correlation for the size of the scale with size -1 / (n - 1), where n = number of items. The resulting value was used as a threshold value for detecting item dependence. With six scale items, it is expected that the average Q_3 is equal to -0.20 if there is no item dependence. Positive correlations are indicative of possible item dependence. For pairs of items with positive correlations, a critical value of 0.20 was used to flag possible item dependency (Chen & Thissen, 1997).

Exhibit A.6 presents the item correlations for the scale at the fourth grade. The average correlation between pairs of items was $\overline{Q_3}$ = -0.19, suggesting that overall, no dependence exists among the items. The highest positive correlation between residuals occurred between items 1C and 2E (Q_3 = 0.09). The second largest positive correlation occurred between items 1A and 1B (Q_3 = 0.07). With both of these values below 0.20, it was determined that no dependency existed among the scale items at the fourth grade.

Exhibit A.6: Autocorrelation Statistics for the Digital Self-Efficacy Scale—Fourth Grade

Items	1A	1B	1C	2A	2B	2E
1A	-					
1B	0.07	-				
1C	-0.20	-0.28	-			
2A	-0.23	-0.21	-0.31	-		
2B	-0.27	-0.27	-0.26	-0.10	-	
2E	-0.36	-0.36	0.09	-0.19	-0.10	-

Exhibit A.7 presents the item correlations for the scale at the eighth grade. At the eighth grade, the average correlation between pairs of items was $\overline{Q_3}$ = -0.20, suggesting that overall, no local dependence existed among the items. The highest positive correlation between item residuals occurred between items 2B and 2E (Q_3 = 0.12). The second largest correlation occurred between items 1C and 2E (Q_3 = 0.04), followed closely by the correlation between items 2A and 2B (Q_3 = 0.03). All three were below the critical value of 0.20. Therefore, no dependency was detected among the six scale items at the eighth grade.

Exhibit A.7: Autocorrelation Statistics for the Digital Self-Efficacy Scale—Eighth Grade

Items	1A	1B	1C	2A	2B	2E
1A	-					
1B	-0.01	-				
1C	-0.18	-0.29	-			
2A	-0.25	-0.25	-0.33	-		
2B	-0.35	-0.31	-0.28	0.03	-	
2E	-0.43	-0.37	0.04	-0.09	0.12	-

Producing Scale Scores

Individual scale scores for each student were produced in Conquest using weighted minimum likelihood estimation (Warm, 1989). Only students with responses to at least two items were given scores. The resulting scale scores were on the logic metric, with values ranging from approximately -5 to 5. A linear transformation was performed at each grade to report the scale scores on the TIMSS reporting metric, with a mean person score of 10 and standard deviation of 2. The linear transformations were applied to the logit scale scores with:

$$\delta_i^* = A + B \cdot \delta_i, \tag{A.2}$$

where δ_i^* is the scale score for person *i* on the TIMSS metric; δ_i is the score for person *i* on the logit metric, and *A* and *B* are linear transformation constants. The linear transformation constants were computed at each grade by:

$$B = \frac{s_{\delta^*}}{s_\delta} \tag{A.3}$$

$$A = M_{\delta^*} - B \cdot M_\delta, \tag{A.4}$$

where s_{δ^*} and s_δ are the standard deviations across the person scores on the target TIMSS metric $(s_{\delta^*} = 2)$ and original logit metric, respectively; and M_{δ^*} and M_δ are the mean person scores on the target $(M_{\delta^*} = 10)$ and original metric, respectively.

Exhibit A.8 shows the resulting scale transformation constants at the fourth and eighth grade.

Exhibit A.8: Scale Transformation Constants for the Digital Self-Efficacy Scale—Fourth and Eighth Grades

	Scale Transformation Constants	
	Fourth Grade	**Eighth Grade**
A	7.30166	6.89635
B	1.64334	1.26552

Validating the Scale

To ensure the scale has at least a small, positive relationship with achievement for the analysis, correlation coefficients were computed for the relationship between the scale scores and both paperTIMSS and eTIMSS achievement. Exhibit A.9 presents the international average correlation coefficients. The relationships with achievement are small at both grades and for both subjects, with correlations around 0.10. The relationships are the same for paperTIMSS and eTIMSS. The strongest relationship with achievement was for fourth grade science scores ($r = 0.13$ for paperTIMSS and eTIMSS). Eighth grade mathematics scores showed the smallest relationship with the scale ($r = 0.05$ for paperTIMSS and eTIMSS).

Exhibit A.9: International Average Correlation between Digital Self-Efficacy and Achievement

Grade/Subject	**International Average Correlation Coefficient (*r*)**	
	paperTIMSS	**eTIMSS**
Fourth Grade (21 countries)		
Mathematics	0.11	0.11
Science	0.13	0.13
Eighth Grade (11 countries)		
Mathematics	0.05	0.05
Science	0.11	0.11

Creating Content-Referenced Regions

To provide a content-referenced interpretation for the resulting scales, the TIMSS and PIRLS method for classifying students into high, middle, and low regions based on their scale scores was implemented (Martin, Mullis, Hooper, Yin, Foy, & Fishbein, 2017). The boundaries of the regions were defined based on judgement in terms of combinations of the response categories. To have "High" digital self-efficacy, students had to respond "Agree a lot" or "I definitely can" to at least three of the six items, and respond "Agree a little" or "I probably again" to the other three, on average. To have "Low" digital self-efficacy, students, on average, would have to respond "Disagree a little" or "I probably can't" to at least three of the six items and "Agree a little" or "I probably can" to the other three. With the raw score points of the items being 3 = "Agree a lot" or "I definitely can," 2 = "Agree a little" or "I probably can," 1 = "Disagree a little" or "I probably cannot," and 0 = "Disagree a lot" or "I definitely cannot," the "High" category would correspond to a raw score of $3 \cdot 3 + 2 \cdot 3 = 15$ and higher. The "Low" category would correspond to a raw score of $2 \cdot 3 + 1 \cdot 3 = 9$ and lower. Students with "Medium" digital self-efficacy would have raw scores between 9 and 15.

All Rasch scales have a unique scale score associated with each possible raw score. Exhibit A.10 shows the range of possible raw scale scores the equivalent scores for the scales at the fourth and eighth grade, respectively. With six scale items at each grade, raw scores between 0 and 18 are possible. Using the values produced by Conquest on the logit metric, the logit scores were transformed to the TIMSS reporting metric at each grade using the scale transformation constants in Exhibit A.8. The resulting cutpoints are the transformed scale score values (rounded to one decimal point) associated with 9 and

15, respectively. The low cutpoint value was rounded up and the high cutpoint value was rounded down to make sure all scale scores are included.

Exhibit A.10: Equivalence Table of Raw and Transformed Scale Scores for the Digital Self-Efficacy Scale—Fourth and Eighth Grades

Raw Score	Fourth Grade		Eighth Grade	
	Transformed Scale Score	Cutpoint	Transformed Scale Score	Cutpoint
0	2.60486		2.61154	
1	4.15694		3.86879	
2	4.86841		4.45648	
3	5.35088		4.86302	
4	5.72944		5.18975	
5	6.05167		5.47593	
6	6.34114		5.74186	
7	6.61211		6.00066	
8	6.87454		6.26252	
9	7.13642	7.2	6.53671	6.6
10	7.40596		6.83311	
11	7.69120		7.16282	
12	8.00286		7.53903	
13	8.35585		7.97813	
14	8.77219		8.50127	
15	9.28764	9.2	9.13612	9.1
16	9.96695		9.91945	
17	10.96054		10.93329	
18	12.94319		12.66245	

Exhibit A.11 shows the resulting percentages of students in each of the regions of digital self-efficacy to examine in preparation for interpreting the results of the analysis in the next section. There were relatively low percentages of students in the "Low" category, with the majority of students having "High" digital self-efficacy (61.9% and 68.1% at the fourth and eighth grades, respectively). This was expected due to the nature of responses to the items shown in Exhibits A.1 and A.2.

Exhibit A.11: Percentages of Students by Level of Digital Self-Efficacy—Fourth and Eighth Grades

Digital Self-Efficacy	International Average Percent of Students	
	Fourth Grade	**Eighth Grade**
Low	6.4%	4.3%
Medium	29.1%	25.2%
High	61.9%	68.1%