Fishbein *et al. Large-scale Assess Educ* (2018) 6:11
https://doi.org/10.1186/s40536-018-0064-z

Large-scale Assessments in Education

RESEARCH Open Access

The TIMSS 2019 Item Equivalence Study: examining mode effects for computer-based assessment and implications for measuring trends

Bethany Fishbein*, Michael O. Martin, Ina V. S. Mullis and Pierre Foy

*Correspondence:
bethany.fishbein@bc.edu
TIMSS & PIRLS International Study Center, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA

Abstract

Background: TIMSS 2019 is the first assessment in the TIMSS transition to a computer-based assessment system, called eTIMSS. The TIMSS 2019 Item Equivalence Study was conducted in advance of the field test in 2017 to examine the potential for mode effects on the psychometric behavior of the TIMSS mathematics and science trend items induced by the change to computer-based administration.

Methods: The study employed a counterbalanced, within-subjects design to investigate the potential for eTIMSS mode effects. Sample sizes for analysis included 16,894 fourth grade students from 24 countries and 9,164 eighth grade students from 11 countries. Following a review of the differences of the trend items in paper and digital formats, item statistics were examined item by item and aggregated by subject for paperTIMSS and eTIMSS. Then, the TIMSS scaling methods were applied to produce achievement scale scores for each mode. These were used to estimate the expected magnitude of the mode effects on student achievement.

Results: The results of the study provide support that the mathematics and science constructs assessed by the trend items were mostly unaffected in the transition to eTIMSS at both grades. However, there was an overall mode effect, where items were more difficult for students in digital formats compared to paper. The effect was larger in mathematics than science.

Conclusions: Because the trend items cannot be expected to be sufficiently equivalent across paperTIMSS and eTIMSS, it was concluded that modifications must be made to the usual item calibration model for TIMSS 2019 to measure trends. Each eTIMSS 2019 trend country will administer paper trend booklets to a nationally representative sample of students, in addition to the usual student sample, to provide a bridge between paperTIMSS and eTIMSS results.

Keywords: Computer-based assessment, Mode effects, TIMSS, Trends over time

Introduction

IEA's TIMSS (the Trends in International Mathematics and Science Study)[1] is an international comparative study of student achievement in mathematics and science at the fourth and eighth grades. Conducted on a four-year assessment cycle since 1995, TIMSS has assessed student achievement using paper-and-pencil methods on six occasions—in 1995, 1999 (eighth grade only), 2003, 2007, 2011, and 2015—and has accumulated 20 years of trend measurements (Martin et al. 2016a; Mullis et al. 2016). Now for the 2019 assessment cycle, TIMSS is transitioning to a computer-based "eAssessment system," called eTIMSS. Just over half of the 65 TIMSS countries are administering eTIMSS in 2019, while the remainder administer TIMSS in paper-and-pencil format, as in previous TIMSS cycles.

The shift from the traditional paper-and-pencil administration to a fully computer-based testing system promises operational efficiencies, enhanced measurement capabilities, and extended coverage of the TIMSS assessment frameworks in mathematics and science. Students from the participating eTIMSS countries will take the assessment on personal computers (PCs) or tablets and will have digital tools available through the eTIMSS interface, including a number pad, ruler, and calculator. eTIMSS 2019 includes extended Problem Solving and Inquiry Tasks (PSIs) and a variety of digitally enhanced item types, including drag and drop, sorting, and drop-down menu input types.

It is acknowledged that changing from paper-and-pencil to the new PC- and tablet-based administration could have substantial and unpredictable effects on student performance (APA 1986; Bennett et al. 2008; Jerrim et al. 2018; Mazzeo and von Davier 2014), which would have to be taken into consideration in analyzing and reporting the TIMSS 2019 results. These "mode effects" 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 et al. 2002; Horkay et al. 2006; Zhang et al. 2016).

The TIMSS 2019 Item Equivalence Study was designed to discover as much as possible about the potential impact of converting from paper-and-pencil to computer-based assessment while the assessment was still under development. Students in the countries participating in the study were asked to respond to TIMSS mathematics and science items in both eTIMSS and paperTIMSS modes of administration, and the results were analyzed for evidence of mode effects. By conducting the study in 2017, before both the field test (2018) and main data collection (2018–2019), the study informed the item development process and provided crucial information about the likely mode effects that will need accounting for in reporting the results of TIMSS 2019. The study was conducted, data analyzed, and results reported over a very short period of time, and utilized straightforward analytic techniques both for efficiency and ease of reporting to a wide audience.

Challenges in digitally transforming an international large-scale assessment

TIMSS will extend its 20 years of trends in mathematics and science achievement in 2019 while transitioning to a digital environment. In measurement terms, TIMSS 2019 faces

[1] TIMSS is directed by the TIMSS & PIRLS International Study Center at Boston College on behalf of IEA, the International Association for the Evaluation of Educational Achievement.

the challenge of continuing trends from the previous TIMSS assessments (1995–2015), which were established with paper-and-pencil methods, while also maintaining comparability with the paper version of TIMSS 2019 (paperTIMSS). The TIMSS approach to measuring trends involves retaining a substantial portion of the items (approximately 60%) from each assessment cycle to be administered in the next cycle. Since these "trend items" are identical for adjacent assessment cycles (e.g., 2011 and 2015), they form the basis for a common item linking of TIMSS achievement scales from cycle to cycle using item response theory (IRT) methods (Foy and Yin 2016).

Because measuring trends in TIMSS assessments prior to 2015 depended on having trend items that were identical from cycle to cycle, it was accepted that the 2019 digital versions of the 2015 trend items should be as similar as possible to the paper versions used in 2015. Keeping the eTIMSS and paperTIMSS versions of the trend items as similar as possible also helps ensure the comparability of the eTIMSS and paperTIMSS versions of the 2019 assessments. Therefore, TIMSS converted the trend items to eTIMSS format with the goal of reducing the potential for mode effects and maintaining equivalence as much as possible.

TIMSS developed the eAssessment system to be compatible with a variety of digital devices to keep up with continuously emerging technologies and allow countries to use existing digital devices as much as possible. In accommodating diversity in digital devices across countries, TIMSS also had to consider the possibility of device effects causing further variation in student performance between modes (Davis et al. 2017; DePascale et al. 2016; Strain-Seymour et al. 2013; Way et al. 2016).

Investigating mode effects

When converted to eTIMSS format, a large proportion of the trend items (about 80%) appeared essentially identical to their paperTIMSS counterparts, while the remainder needed modification to some extent. Having so many apparently identical trend items raised the possibility that many of them would show no mode effect, and would have equivalent psychometric properties regardless of the mode of administration—eTIMSS or paperTIMSS. The TIMSS 2019 Item Equivalence Study was designed as a controlled experiment to test this proposition in advance of the TIMSS 2019 field test and main data collection. In addition to testing for possible mode effects, the Item Equivalence Study provided an ideal opportunity to try out the eTIMSS user interface and other components of the eAssessment system in a realistic classroom environment with a variety of digital devices. It was administered in April and May of 2017 and the data were analyzed through December 2017.

To examine the mode effect, trend items were administered to samples of students in participating countries according to a counterbalanced design, with half the students taking items in eTIMSS format first and then items in paper format, and the other half taking paperTIMSS items first and eTIMSS items second. Each student was administered the full TIMSS experience, two blocks of mathematics and two blocks of science items, in each format, care being taken to ensure a different selection of items in each administration. With fairly large samples (800 students) and the entire pool of trend items administered in each country, the Item Equivalence Study was well positioned

Table 1 List of participating countries

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

both to detect possible mode effects and to estimate the impact of such effects on the measurement of trends.

The common item linking methodology that TIMSS uses to maintain comparability of results between assessment cycles expects that the "common" items mostly behave the same from cycle to cycle. However, the new eTIMSS mode of administration has the potential to change the psychometric properties of the trend items to the extent that a different approach to linking paperTIMSS and eTIMSS results may be necessary to preserve trends. To inform the methods and procedures necessary to ensure results comparable to TIMSS 2015, the TIMSS 2019 Item Equivalence Study addressed the following research questions:

(1) To what extent can the eTIMSS and paperTIMSS versions of the TIMSS 2019 trend items be considered psychometrically equivalent, and hence usable in a common item linking design?
(2) To the extent that the eTIMSS and paperTIMSS versions are not equivalent, what would be the likely impact of this mode effect on the measurement of trends in TIMSS 2019?

Methods

The first research question was addressed by (a) comparing the eTIMSS and paperTIMSS versions of each of the trend items with a view to identifying areas of difference that could contribute to a mode effect, and (b) conducting an item-by-item analysis of performance differences in paper and digital formats. This analysis examined differences in item difficulty (percent correct), item discrimination (point-biserial correlations), percent omitted, and percent "not reached."

Examining the likely impact of a mode effect on the measurement of trends (the second research question) involved estimating student proficiency on the TIMSS achievement scales using the usual TIMSS scaling methodology—IRT scaling combined with latent regression. Having proficiency scores for each student in both paperTIMSS and eTIMSS modes enabled (a) estimating mode effects on TIMSS achievement scales

Table 2 Booklet/item block combination design—fourth and eighth grades

Paper booklet	Item 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

overall (mathematics and science at both fourth and eighth grades) and (b) estimating mode effects for student subgroups.

Sample

Twenty-five countries participated in the TIMSS 2019 Item Equivalence Study, with 24 countries at the fourth grade and 13 countries at the eighth grade (Table 1). Each country was responsible for selecting a purposive sample of 800 students at each grade that included students with a range of abilities and backgrounds. For example, participating schools and classes should ideally have included both low- and high-achieving students. Some countries had better success in achieving a diverse school sample than others. For analysis purposes, each student was assigned a sampling weight of 1.

Instruments

The complete set of mathematics and science trend items brought forward from TIMSS 2015 was administered for the study—187 items at the fourth grade and 232 items at the eighth grade. For each grade and subject, the items were distributed among eight item blocks, each mimicking the distribution of item types as well as the content and cognitive skills that the entire assessment is meant to cover. Approximately half the score points came from multiple-choice items and the other half from constructed response items.

Under a counterbalanced design, each student sampled for the study received a full paper-and-pencil booklet (paperTIMSS) of trend items and an equivalent set of trend items as an eTIMSS "item block combination." Students also completed a questionnaire addressing student characteristics, including gender, socioeconomic status, and their attitudes for using computers and tablets. Participating countries administered eTIMSS on PC or Android tablet devices.

For both the fourth and eighth grades, the trend item blocks were assigned to eight paper booklets and eight equivalent eTIMSS item block combinations (Table 2). Of the eight booklets, six were identical to booklets administered in 2015. The other two booklets contained two trend item blocks that were also paired together in 2015, but with

Table 3 Counterbalanced booklet rotation design—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

two other blocks (M04 and S04) repositioned to replace blocks that were removed after the 2015 cycle. Consistent with the matrix sampling design used for each TIMSS cycle, each item block appears in two booklets in different positions to provide a mechanism for linking student responses across booklets for scaling. Each booklet 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 the booklets begin with two blocks of mathematics items and half begin with two blocks of science items. This item block distribution scheme was replicated for the eight eTIMSS item block combinations (Table 2).

Counterbalanced research design

Each student was assigned one paperTIMSS test booklet and one eTIMSS item block combination according to a counterbalanced rotation scheme (Table 3). In each country, half of the students were assigned paperTIMSS first (Booklets 1–8), and half of the students were assigned eTIMSS first (Item Block Combinations ET19PTBC01–08). The rotation scheme ensured that students did not encounter the same items in paperTIMSS and eTIMSS format. For example, students assigned Booklet 1 were assigned eTIMSS Item Block Combination ET19PTBC03, and students assigned Booklet 2 were assigned eTIMSS Item Block Combination ET19PTBC04. These booklets/item block combinations had no items in common (see Table 2). The two test sessions occurred either within the same day or across two consecutive days. Student Tracking Forms were used by each participating school to ensure proper implementation of the counterbalanced design.

Results

To address the first research question—the extent to which the paperTIMSS and eTIMSS versions of the TIMSS 2019 trend items can be considered psychometrically equivalent—the study began with a comparison of the eTIMSS and paperTIMSS versions of each trend item. This involved developing a set of explicit criteria for classifying the items according to their differences across paper and digital formats as well as characteristics that may be expected to induce mode effects. Staff from the TIMSS & PIRLS International Study Center classified the trend items according to their hypothesized likelihood for being "strongly equivalent" or "invariant" between paperTIMSS and eTIMSS.

Preliminary item classification descriptions were developed based on the results of earlier small-scale pilot studies and the mode effect literature relevant to the types of items in the study. The following types of items or features of items were of particular interest:

- Differences in presentation between paper and digital formats (Pommerich 2004), such as formatting changes necessary to render the item on a digital interface (Sandene et al. 2005).
- Complex graphs or diagrams (Mazzeo and Harvey 1988) or heavy reading, possibly requiring greater cognitive processing (Chen et al. 2014; Noyes and Garland 2008).
- Scrolling required to view all parts of the item (Bridgeman et al. 2003; Pommerich 2004; Way et al. 2016).
- 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 human-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, requiring students to transcribe calculations from scratch paper to the PC or tablet (Johnson and Green 2006).
- Items with numerical answers requiring the "number pad" to input the response.
- Items requiring the use of the "drawing" feature to draw or label features (Sandene et al. 2005; Strain-Seymour et al. 2013).

eTIMSS constructed response items had one of three different "input types," specifying the action required by students to respond. The "keyboard" input required students to use the full keyboard equipped by the delivery device (either external or on-screen) to type responses, including mathematical equations. For numerical responses, students used the on-screen "number pad" with digits 0 through 9, a decimal point, negative sign, and division symbol. "Drawing" input types required students to show work, draw, or label diagrams.

Two raters refined the pre-developed criteria into detailed descriptions and used them to classify each item into one of four types—"Identical," "Nearly Identical," "Worrisome," or "Severe" (see full criteria in Fishbein 2018). The 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 was required for a correct response. When the two raters disagreed, a third rater who was also familiar with the trend items made the final classification.

The results indicated that the majority of the trend items at each grade were considered to be "Identical" or "Nearly Identical"—assessing exactly the same construct and maintaining their presentation in both modes. These items could reasonably be expected a priori to perform the same for both paperTIMSS and eTIMSS. A larger proportion of eighth grade items was classified as "Identical" or "Nearly Identical" compared to fourth grade items, under the a priori assumption that eighth grade students are more familiar with using digital devices.

With number pad and drawing input types being much more common in mathematics, a larger proportion of mathematics items compared to science items was classified as "Worrisome" or "Severe," hypothesized to behave differently for eTIMSS. Unfortunately, the number pad input feature was not completely functional at the time of the study. At the fourth grade in particular, inputting numbers such as fractions was cumbersome for students. Additionally, the results of earlier pilot studies indicated that students found the drawing feature difficult to use, and the scoring system was unable to reproduce students' responses to these items for scoring at the time of the study. The results of subsequent item-by-item analysis found that most of the data were lost for these items—13 items at the fourth grade and 11 items at the eighth grade.

Despite the difficulties described above, the comparison of the eTIMSS and paperTIMSS versions indicated that efforts to keep the trend items looking the same and maintain construct equivalence across paper and digital formats were mostly successful. Following this review, staff at the TIMSS & PIRLS International Study Center further refined the classifications for the item-by-item analysis.

Item-by-item analysis of performance differences

Having completed the classification of items by likely degree of mode effect based on the appearance of the items, the next step in addressing the first research question was to examine item equivalence in terms of student performance. This was done by comparing descriptive item statistics for each item based on the two administration modes. For each item included in the TIMSS 2019 Item Equivalence Study—92 mathematics items and 95 science items at the fourth grade, and 114 mathematics items and 118 science items at the eighth grade—the percent correct, point-biserial correlation, percent omitted, and percent "not reached" were calculated for both paperTIMSS and eTIMSS data. "Difference statistics" were produced for each item by subtracting the eTIMSS statistic from the paperTIMSS statistic (e.g., $p_{paper} - p_{eTIMSS}$). This produced indicators of the mode effect for each item by country.

The review of item statistics included reexamining the previous item classifications and identifying any "Worrisome" or "Severe" items that may not be suitable for the eTIMSS environment, due to the inability of students in the study to appropriately respond to the item as they would on paper or other limitations of the eAssessment system at the time. Items having a difference of at least 10% in omitted responses between paperTIMSS and eTIMSS in five or more countries were checked against the countries' item-by-item documentation for reported differences between paper and digital versions of the items.

The review of the item difference data 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) whose eTIMSS versions were clearly not equivalent to their paper versions. These items were re-classified as *expected non-invariant*. The remaining items were classified as *expected invariant*, and could reasonably be expected to be psychometrically equivalent.

Table 4 Number of items and sample sizes for analysis

Grade	Total cases	Mathematics		Science	
		Valid items	Average responses per item	Valid items	Average responses per item
Fourth grade (21 countries)	16,894	77	4199	82	4200
Eighth grade (11 countries)	9164	97	2278	110	2270

Counts reflect resulting sample sizes after deleting problematic data and *expected non-invariant* items

Mathematics items fitting the criteria for *expected non-invariant* mostly included those with drawing inputs that could not be scored effectively at the time of study and items with fraction answers that could not be input due to limitations of the number pad. In science, *expected non-invariant* items included those with keyboard entry boxes in tables that were too restrictive for character-based languages. A few items with severe scrolling had omit rates as high as 50% for eTIMSS, substantially higher than omit rates for paperTIMSS. Countries' reports suggest that students may not have seen some parts of items that required substantial scrolling to uncover.

Table 4 presents the number of items and final sample sizes for analysis after eliminating *expected non-invariant* items. Data for two countries at each grade were excluded from analysis because of technical problems with assessment delivery devices and issues with the data upload server.

The final database was used to compute an international average for each of the item statistics for paperTIMSS, eTIMSS, and their differences, respectively, separately by subject and grade. Each country was weighted equally in computing the international average for item difficulty (mean percent correct), item discrimination (mean point-biserial correlation), mean percent omitted, and mean percent not reached. To provide a comparison of the degree that the distributions of item difficulty and discrimination varied for paperTIMSS and eTIMSS, standard deviations for each of these statistics were computed across the items for each country, then pooled across countries. 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.

Items with similar measurement properties across modes should have very similar percent correct values and show only small, random deviations from the identity line when plotted against one another. Clearly this was not the case for the *expected non-invariant* items, where the average percentage of students answering correctly was higher for paperTIMSS than for eTIMSS in almost every instance, indicating that the items were more difficult in eTIMSS than in paperTIMSS (see Figs. 1, 2). As expected, these items showed definite evidence of a mode effect.

The results for the *expected invariant* items were more encouraging, with most items clustering close to the identity line in plots for both subjects and grades. Upon closer inspection, however, the results provide evidence of a general mode effect for the TIMSS trend items. Particularly for mathematics (Fig. 1), most points clustered just below the identity line at each grade, indicating the items generally were more difficult for eTIMSS than for paperTIMSS. For science (Fig. 2), there was more equal distribution of points

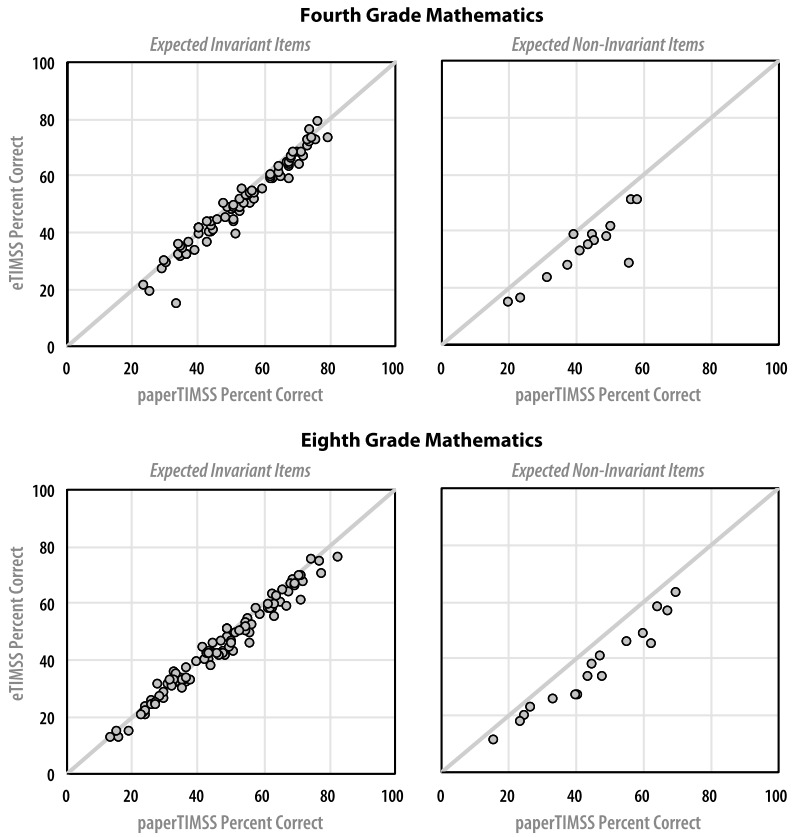

Fig. 1 Item plots of international average percent correct statistics for mathematics—paperTIMSS vs. eTIMSS. Each data point represents a mathematics trend item with *x*-axis the percent correct based on paperTIMSS data and *y*-axis percent correct based on eTIMSS data. *Expected invariant* items (left) were included in the analysis. *Expected non-invariant* items (right) were plotted as a means of comparison

around the identity line, suggesting the mode effect for science may be smaller than for mathematics.

Further averaging the international percent correct across all of the *expected invariant* items revealed the mode effect more clearly (Table 5). Fourth grade mathematics items showed the largest average difference in item difficulty between paperTIMSS and eTIMSS, with an average difference between modes of 3.6 percentage points. Eighth grade mathematics items showed a similar effect, with a 3.4 percentage-point difference between paperTIMSS and eTIMSS, on average. Science items at both grades showed smaller mode effects on item difficulty compared to mathematics items, with average differences of 1.7 percentage points at the fourth grade and 1.5 percentage points at the eighth grade.

The size of the average point-biserial correlations suggest there was little or no effect of mode of administration on item discrimination statistics for the *expected invariant* items, 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 (Table 6). Similarly, after removing the *expected non-invariant* items, percentages of missing responses—both omitted and not reached—were practically identical for paperTIMSS and eTIMSS (Fishbein 2018). At the fourth grade, mathematics items had approximately

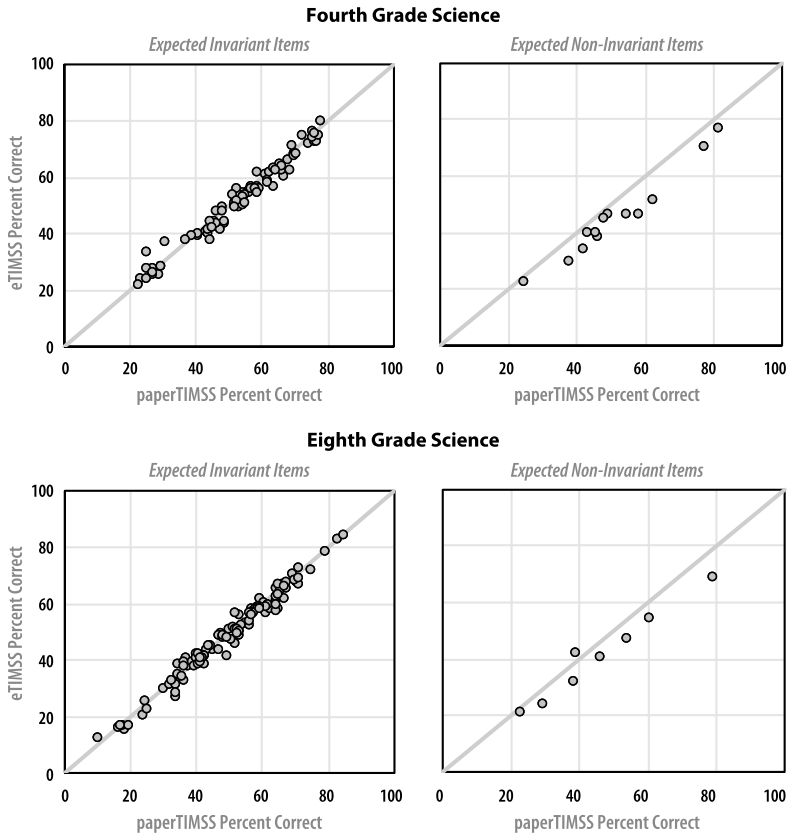

Fig. 2 Item plots of international average percent correct statistics for science—paperTIMSS vs. eTIMSS. Each data point represents a science trend item with *x*-axis the percent correct based on paperTIMSS data and *y*-axis percent correct based on eTIMSS data. *Expected invariant* items (left) were included in the analysis. *Expected non-invariant* items (right) were plotted as a means of comparison

Table 5 Item percent correct, averaged across countries and across items

Grade/subject	Average percent correct across countries and items		
	paperTIMSS	eTIMSS	Difference
Fourth grade			
Mathematics	53.7 (18.1)	50.1 (18.0)	3.6 (6.0)
Science	53.1 (17.9)	51.3 (17.5)	1.7 (6.0)
Eighth grade			
Mathematics	47.4 (19.1)	44.0 (18.5)	3.4 (5.3)
Science	49.6 (18.4)	48.1 (18.5)	1.5 (5.7)

( ) Standard deviations appear in parentheses. Standard deviations of item percent correct were computed across items in each country and then pooled across countries. Because of rounding some results may appear inconsistent

5.1% of responses missing for paperTIMSS and 5.8% missing for eTIMSS, on average. Fourth grade science had 6.0% of responses missing on average across paperTIMSS items and 6.3% missing on average across eTIMSS items. At the eighth grade, approximately 4.9% of mathematics item responses were missing for paperTIMSS with 5.1% of responses missing for eTIMSS. In science, items had an average of 4.9% missing and

Table 6 Item point-biserial correlations, averaged across countries and across items

Grade/subject	Average point-biserial correlation across countries and items		
	paperTIMSS	eTIMSS	Difference
Fourth grade			
Mathematics	0.42 (0.12)	0.41 (0.12)	0.01 (0.09)
Science	0.37 (0.10)	0.36 (0.11)	0.00 (0.09)
Eighth grade			
Mathematics	0.42 (0.13)	0.41 (0.14)	0.02 (0.03)
Science	0.37 (0.12)	0.37 (0.12)	0.01 (0.09)

( ) Standard deviations appear in parentheses. Standard deviations of item point-biserial correlations were computed across items in each country and then pooled across countries. Because of rounding some results may appear inconsistent

5.1% missing for paperTIMSS and eTIMSS, respectively. These results suggest there was no effect of eTIMSS on the TIMSS mathematics and science constructs measured by the paper instruments—only item difficulties showed differences (Winter 2010).

Estimating mode effects on TIMSS achievement scales

Given that the item-by-item analyses showed evidence of a general mode effect, with most items, particularly in mathematics, exhibiting an effect to some degree, it was decided to move on to the second research question focusing on the likely impact of a mode effect on the measurement of trends. This analysis first examined the overall mode effect for mathematics and science scores before moving on to a consideration of differential mode effects for student subgroups. For these analyses, it was necessary to derive estimates of student proficiency by applying the TIMSS IRT scaling methodology (Martin et al. 2016b; Mislevy 1991) to the Item Equivalence Study data for both eTIMSS items and paperTIMSS items.

Because the main concern of the study was the effect of moving from paper-based items to computer-based items, the paperTIMSS data was chosen as the baseline against which to compare the eTIMSS data. Therefore, in scaling the data, item parameters were first estimated for the paperTIMSS items, 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 provides an estimate of the expected mode effect size if nothing is done to control for it.

Following the usual TIMSS procedures for scaling the achievement item data (Foy and Yin 2016), item parameters were estimated using mixed IRT models (two- and three-parameter and generalized partial credit), with each country's response data contributing equally to a single, overall calibration. Item calibration was conducted separately by grade and subject using PARSCALE software (Muraki and Bock 1991).

To produce accurate achievement estimates for populations and subpopulations of students with matrix sampling of items, TIMSS uses latent regression with plausible values methodology (Martin et al. 2016b; 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

Table 7 International average scale scores, standard deviations, standard errors, and cross-mode correlation coefficients

Grade/subject	International average scale score									Cross-mode correlation coefficient (r_{adj})
	paperTIMSS			eTIMSS			Difference			
	M	*SD*	*SE*	*M*	*SD*	*SE*	*M*	*SD*	*SE*	
Fourth grade										
Mathematics	527	86	1.4	513	86	1.4	14	40	0.7	0.96
Science	526	86	1.5	518	86	1.5	8	41	0.6	0.96
Eighth grade										
Mathematics	511	98	2.2	497	97	2.3	14	42	1.0	0.97
Science	521	92	2.6	514	90	2.5	7	43	1.0	0.96

Because of rounding some results may appear inconsistent

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.

DGROUP software (Rogers et al. 2006) was used to estimate student proficiency, separately by grade for paperTIMSS and eTIMSS, respectively. Mathematics and science proficiencies were estimated concurrently for each student using a two-dimensional latent regression model. Conditioning variables in the latent regression included questions 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 mean achievement (based on *expected a posterior* scores from PARSCALE), country, and the interactions between country and each other conditioning variable.

The resulting distributions of plausible values were transformed to an approximate TIMSS scale, as follows. First, a Stocking-Lord transformation (Stocking and Lord 1983) was applied, using the TIMSS 2015 IRT item parameters (Foy 2017) to place the resulting plausible values on the TIMSS 2015 theta scales. Then, the same linear transformation constants that were used to transform the TIMSS 2015 theta scores onto the TIMSS reporting metric were applied (see Foy and Yin 2016), resulting in student proficiency scores on the TIMSS reporting scale.

The plausible values were used to produce evidence of construct equivalence and score comparability between paperTIMSS and eTIMSS. Commonly accepted criteria for 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).

Analyses were conducted separately for the fourth grade and eighth grade for mathematics and science, respectively. For each grade and subject, international average scale scores, standard deviations, and standard errors were computed for both paperTIMSS and eTIMSS. International average difference scores were produced by subtracting each eTIMSS plausible value from its corresponding paperTIMSS plausible value for each case and averaging the results. To treat each country equally in analyses across countries, each country's sample was weighted to give a sample size of 500 students.

The results were produced using IEA's IDB Analyzer software and IBM SPSS Statistics Software Version 24. For each analysis, IEA's IDB Analyzer software applied sampling weights, computed the average of each plausible value across all cases in the database, and aggregated the results across the plausible values for interpretation. It also produced standard errors for each using the jackknife repeated replication method (Foy and LaRoche 2016; Rust 2014). Because the student samples were not drawn randomly, the standard errors are not an accurate reflection of the population data. However, they reflect the variance between schools as well as the imputation error, and are a useful indicator of the variability of the Item Equivalence Study data.

Average scale scores based on paperTIMSS data were higher than scores based on eTIMSS data, confirming that the trend items were more difficult under the conditions of the eTIMSS delivery (Table 7). As expected from the item analyses, the effect was larger for mathematics than for science. At the fourth and eighth grades, there was an average difference across countries of 14 points for mathematics scores. Science showed an average difference of 8 score points at the fourth grade and 7 score points at the eighth grade. Standard deviations and standard errors were approximately equal across modes.

In the TIMSS context, a difference of 14 points in mathematics scores is substantial and corresponds to one-fourth 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 et al. 1998; Mullis et al. 1998). These international average results from TIMSS 1995 are similar to more recent results from TIMSS 2015, when Norway participated with two grade levels of students taking the fourth and eighth grade assessments, respectively (Martin et al. 2016a; Mullis et al. 2016). 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 sampling requirements are designed to yield a standard error no greater than 3.5% of the standard deviation associated with each country's mean achievement score (LaRoche et al. 2016). A standard deviation corresponds to approximately 100 points on the TIMSS reporting scale, so student samples should provide for a standard error of 3.5 points. This corresponds to a 95% confidence interval of ± 7 score points for an achievement mean and ± 10 score points for the difference between means from adjacent assessment cycles. Therefore, a 14-point difference would constitute a substantial difference between mean scores and must be taken into account in linking eTIMSS to the TIMSS achievement scale.

The cross-mode (eTIMSS-paperTIMSS) correlations were very large ($r > 0.95$) for each grade and subject (Table 7), suggesting that despite the differences in mean achievement, students' proficiency scores ranked similarly in both modes and that eTIMSS did not have an effect on the TIMSS mathematics and science constructs. Examination of mean scores by country confirmed 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 by a negligible amount toward the middle. However, the large standard

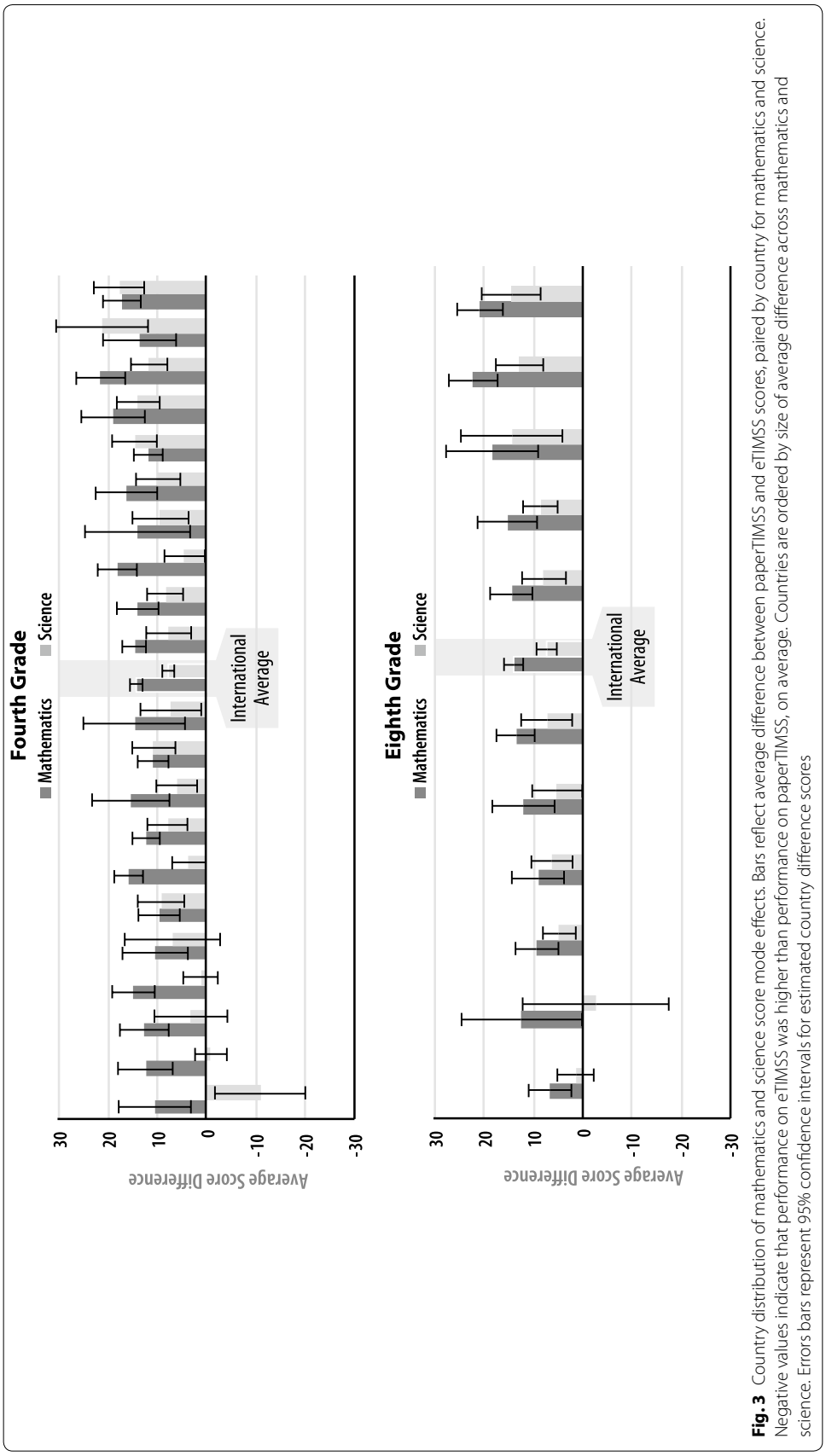

Fig. 3 Country distribution of mathematics and science score mode effects. Bars reflect average difference between paperTIMSS and eTIMSS scores, paired by country for mathematics and science. Negative values indicate that performance on eTIMSS was higher than performance on paperTIMSS, on average. Countries are ordered by size of average difference across mathematics and science. Errors bars represent 95% confidence intervals for estimated country difference scores

deviations reported for the difference values (Table 7) suggest that the magnitude of mode effects differed substantially across students.

Although interpretation of country-level results is not possible without nationally representative samples, the size of the 95% confidence intervals around each country mean difference score suggests that much of the difference in mode effect across countries may be due to sampling error, at least for these data. However, there was some variation in the mathematics and science mode effects across countries (Fig. 3). There was more variation in science mode effects compared to mathematics, despite the mathematics score differences being larger.

Estimating mode effects for student subgroups

Addressing the second part of the second research question, the final series of analyses examined differences between paperTIMSS and eTIMSS proficiency scores in relation to student background variables. The results provided additional information about the equivalence of the mathematics and science constructs between modes (Randall et al. 2012). If two scores are measuring the same construct, then they should have the same degree of relationship with other related measures (APA 1986; DePascale et al. 2016; Winter 2010).

The analysis was conducted using three grouping variables identified in the literature to relate to mode effects:

- Socioeconomic status (Bennett et al. 2008; Jerrim 2016; MacCann 2006; Zhang et al. 2016).
- Gender (Cooper 2006; Gallagher et al. 2002; Jerrim 2016; Parshall and Kromrey 1993).
- Confidence in using computers and tablets, or "digital self-efficacy" (Cooper 2006; Pruet et al. 2016; Zhang et al. 2016).

All subgroup analyses were conducted separately for the fourth and eighth grades and for mathematics and science, respectively, with each country contributing equally to the results. From the student questionnaire, the "Books in the Home" variable was used as a proxy measure of socioeconomic status, which historically has shown to be a strong predictor of achievement in TIMSS (e.g., Mullis et al. 2016; Mullis et al. 2017). Gender data were collected from participating schools via the Student Tracking Forms used for test administration or from students via the questionnaire. For a measure of digital self-efficacy, a one-parameter IRT scale was constructed for each grade based on six questionnaire items asking about students' confidence in using computers and tablets (see scale construction details in Fishbein 2018). A benchmarking procedure was used to classify students' scores into meaningful "Low," "Medium," and "High" categories of digital self-efficacy for a categorical form of the continuous scale variable.

Analysis of paperTIMSS-eTIMSS difference scores and their standard errors by student subgroups revealed that, on average, the mode effect was mostly uniform across student subgroups by Books in the Home, gender, and digital self-efficacy (see Tables 8, 9). The variation in mean difference scores across subgroups was within the margin of error with 95% confidence.

Table 8 Average paperTIMSS-eTIMSS difference scores by student subgroups—fourth grade

Subgroup	Valid cases	Average percent of students	Average difference score	
			Mathematics	Science
Books in the home				
0–10 books	1947	12 (0.4)	12 (1.4)	10 (1.7)
11–25 books	4180	26 (0.5)	13 (0.8)	7 (0.9)
26–100 books	5411	33 (0.4)	15 (0.7)	9 (0.7)
101–200 books	2811	17 (0.4)	16 (0.9)	7 (0.9)
More than 200 books	2201	13 (0.3)	15 (1.1)	8 (1.0)
Gender				
Girls	8627	52 (3.0)	15 (0.8)	7 (0.6)
Boys	8142	49 (3.0)	13 (0.8)	9 (0.8)
Digital self-efficacy				
Low	1094	7 (0.3)	17 (1.7)	11 (2.0)
Medium	4926	30 (0.4)	14 (0.9)	7 (0.9)
High	10,455	63 (0.6)	14 (0.7)	8 (0.6)

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

Table 9 Average paperTIMSS-eTIMSS difference scores by student subgroups—eighth grade

Subgroup	Valid cases	Average percent of students	Average difference score	
			Mathematics	Science
Books in the home				
0–10 books	1241	14 (0.5)	14 (1.7)	9 (1.7)
11–25 books	2160	25 (0.6)	12 (1.4)	7 (1.3)
26–100 books	2670	30 (0.5)	13 (1.1)	6 (1.2)
101–200 books	1450	16 (0.5)	15 (1.3)	4 (1.3)
More than 200 books	1454	16 (0.6)	16 (1.7)	9 (1.7)
Gender				
Girls	4794	53 (1.4)	14 (1.0)	9 (0.9)
Boys	4308	48 (1.4)	13 (1.5)	6 (1.6)
Digital self-efficacy				
Low	379	4 (0.3)	15 (3.1)	8 (3.3)
Medium	2302	26 (0.5)	15 (1.4)	8 (1.3)
High	6286	70 (0.6)	13 (1.1)	6 (1.0)

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

Following the analysis of average differences by subgroup, a repeated measures analysis of variance (ANOVA) was conducted with mode of administration as the within-subjects factor plus three between-subjects factors: Books in the Home (5 levels), gender (2 levels), and digital self-efficacy (3 levels). Following the usual procedures for analyzing plausible values, each full factorial model (one for each grade/subject) was run in SPSS five times—once for each pair of plausible values as the within-subjects variables (paperTIMSS and eTIMSS)—and the results were aggregated for interpretation.

At the fourth grade, the ANOVA models found a significant effect of Books in the Home × mode on mathematics achievement, $F(4, 15{,}018) = 6.22$, $p < 0.001$, $\eta_p^2 = 0.002$,

and a significant effect of gender × mode on science achievement, $F(1, 15{,}018) = 5.77$, $p < 0.05$, $\eta_p^2 < 0.001$. The eighth grade models also found a significant effect of Books in the Home × mode on mathematics achievement, $F(4, 6849) = 2.70$, $p < 0.05$, $\eta_p^2 = 0.002$. However, all three significant effects were very small, accounting for less than 1% of the variance in achievement between modes overall.

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 difference scores, a multiple linear regression analysis was conducted. The outcome variable was the set of plausible values for the difference scores between paperTIMSS and eTIMSS (*PVDIFF*). The following model was specified for each grade and subject:

$$(PVDIFF)_{ij} = B_0 + B_1(DSE)_{ij} + B_2(Books_1)_{ij} + \cdots + B_5(Books_4)_{ij} + B_6(Gender)_{ij} + \varepsilon_{ij},$$

where digital self-efficacy (*DSE*) was a continuous predictor variable; Books in the Home was a dummy-coded predictor variable where "0–10 books" was the reference category and 1 = "11–25 books" ($Books_1$), 2 = "26–100 books" ($Books_2$), 3 = "101–200 books" ($Books_3$), and 4 = "More than 200 books" ($Books_4$); and gender was a dummy-coded predictor variable where "Girls" were the reference group and 1 = "Boys."

The results of the multiple regression analysis corroborated the ANOVA results, but also found a significant effect of Books in the Home on the size of the science mode effects at both grades ($p < 0.05$). However, further analysis showed no clear relationship between this variable and mathematics or science mode effects. Moreover, the predictor variables explained a very small percentage of variance in the mode effects, overall. At the fourth grade, the predictor variables accounted for less than 2% of the variance in mathematics difference scores ($R^2 = 0.015$) and less than 2% of the variance in science difference scores ($R^2 = 0.016$). At the eighth grade, the variables accounted for approximately 1% of the variance in mathematics difference scores ($R^2 = 0.013$) and less than 2% of the variance in science difference scores ($R^2 = 0.016$).

Discussion

The results of the Item Equivalence Study show clear evidence that TIMSS 2019 trend items presented in eTIMSS format were more difficult on average than the paperTIMSS version, especially for mathematics, and this difference needs to be taken into account when linking eTIMSS 2019 to the TIMSS achievement scale. However, the study results also suggest that the measurement of the TIMSS mathematics and science constructs themselves were relatively unaffected by the transition to eTIMSS. The preliminary item review supported the view that the majority of the trend items appeared equivalent in paper and eTIMSS formats, confirming that efforts to convert the paper trend items to eTIMSS were largely successful. The item analysis also found negligible differences in item discrimination statistics between paperTIMSS and eTIMSS.

Despite differences in means due to the mode effect, score-level standard deviations and standard errors were similar across modes and cross-mode correlation coefficients reflecting the relationships between paperTIMSS and eTIMSS scores were large ($r > 0.95$), reflecting similar score distribution shapes and similar ranking of students. Examining mean scores by country for each grade and subject confirmed that country

rankings were about the same for paperTIMSS and eTIMSS. Lastly, the results of the analysis by student subgroups showed 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.

The above findings meet criteria for evidence that the mathematics and science constructs were unchanged in eTIMSS (APA 1986; DePascale et al. 2016; Randall et al. 2012; Winter 2010). Therefore, the difference in scores that resulted from the mode effects can be accounted for through appropriate linking procedures and the paperTIMSS and eTIMSS scores can be put on a common scale.

Implications for measuring trends in TIMSS 2019

Although quite a large-scale study in terms of the number of countries and students involved and the amount of data collected, the Item Equivalence Study was intended to give only a preliminary indication of mode effects when countries participating in TIMSS 2019 could choose between paper-based (paperTIMSS) and computer-based (eTIMSS) versions of the assessment. Given the mode effects found by the study, it is considered unlikely that the trend items in paperTIMSS and eTIMSS overall will be sufficiently equivalent for the common item linking usually implemented by TIMSS for measuring trends, and that the procedure should be augmented by an additional data source. Accordingly, in addition to administering the full eTIMSS assessment to the usual national sample of about 4500 students, each eTIMSS country will administer the paper trend items to a matched sample of 1500 students (known as the "bridge" sample), resulting in randomly equivalent student samples taking both eTIMSS and paperTIMSS items in each eTIMSS country.

The bridge data will provide a secure basis for linking paperTIMSS and eTIMSS in 2019, regardless of whether any of the trend items can be considered psychometrically equivalent. However, because of improvements made to many of the eTIMSS trend items as well as to the usability and reliability of the eTIMSS delivery platform, there are grounds to expect that individual item mode effects may be less apparent in the data from the main data collection. Because of this, the linking procedure will come after a reexamination of paperTIMSS-eTIMSS data for any evidence of bias due to mode effects. Depending on the outcome of this item analysis, it is possible that a subset of trend items may be identified that are invariant across paperTIMSS and eTIMSS and can be considered common items for calibration purposes.

The approach for measuring trends in TIMSS 2019 (Fig. 4) involves one overall concurrent item calibration to estimate both paperTIMSS and eTIMSS item parameters, based on all assessment data from TIMSS 2015 and TIMSS 2019, and two separate linear transformations. If there are any invariant trend items, these item parameters will be fixed to be equal for paperTIMSS and eTIMSS, and non-invariant eTIMSS item parameters will be estimated freely.

The first linear transformation will place the TIMSS 2019 data from paperTIMSS countries and the paper-based trend item data from the eTIMSS bridge sample on the TIMSS scale by aligning the TIMSS 2015 data under the 2019 concurrent calibration

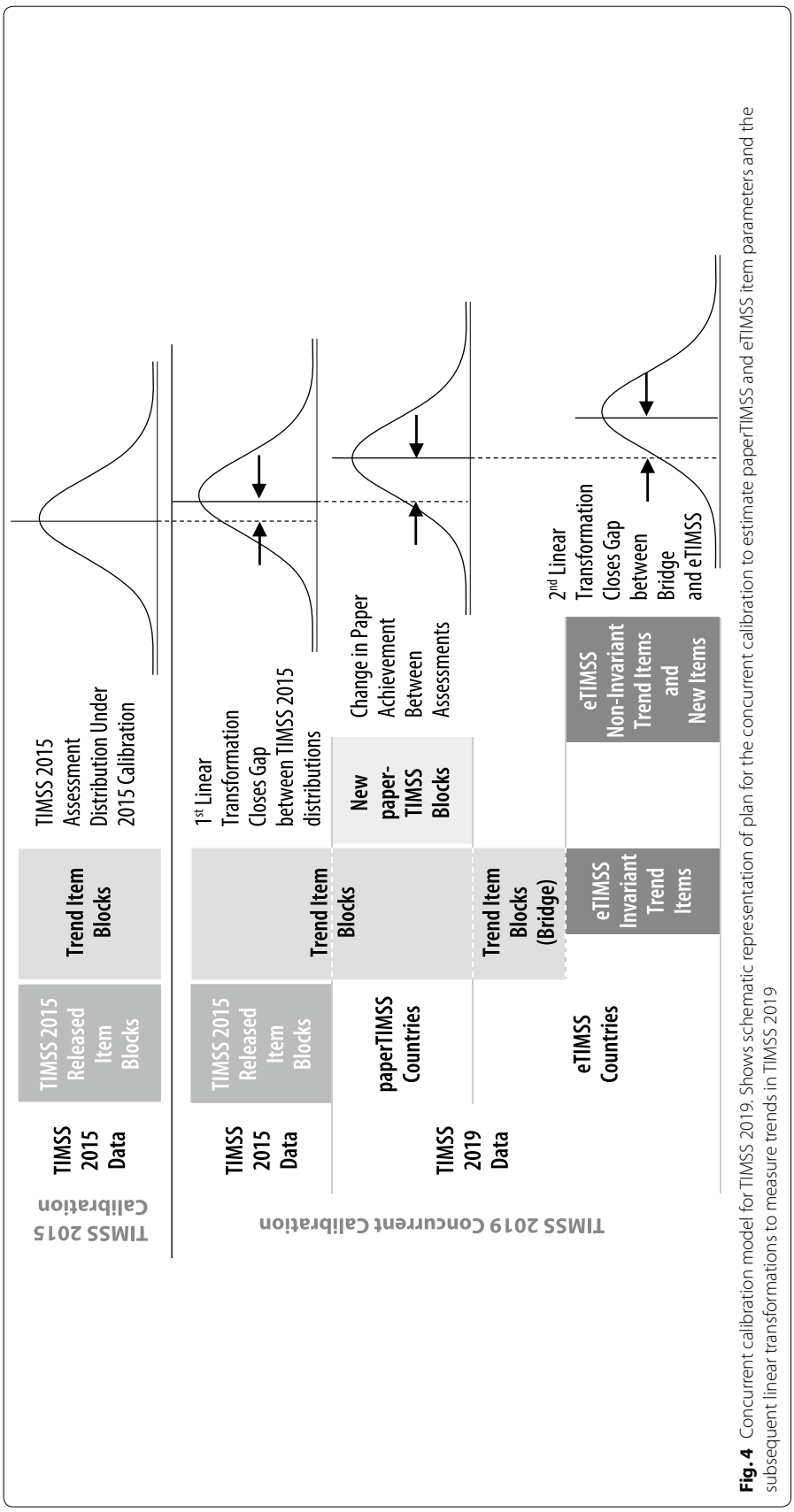

Fig. 4 Concurrent calibration model for TIMSS 2019. Shows schematic representation of plan for the concurrent calibration to estimate paperTIMSS and eTIMSS item parameters and the subsequent linear transformations to measure trends in TIMSS 2019

with the same data under the 2015 calibration. After applying the linear transformation to the TIMSS 2019 scores through common item linking, the score differences that remain between assessments will reflect the change in student achievement over time.

The second linear transformation for eTIMSS 2019 countries will align the distribution of the eTIMSS scores with the already transformed distribution of the paper-based bridge scores through equivalent groups or common population linking, which is possible because the eTIMSS data and the bridge data are equivalent samples from each country's student population. Then, the eTIMSS 2019 scores will be directly comparable with paperTIMSS 2019 scores, as well as TIMSS scores from all previous assessments. This two-step procedure is analogous to the procedure used in TIMSS 2007 to link the TIMSS achievement scales despite a major change in booklet design from 2003 to 2007 (Foy et al. 2008).

Conclusion

The TIMSS 2019 Item Equivalence Study played a valuable role in predicting the likely existence of a mode effect in the main TIMSS 2019 data collection and confirming the need to add a paper-based bridge to the data collection design. This enhanced design, incorporating both eTIMSS and paperTIMSS data, ensures that the measurement of trends will be safeguarded as TIMSS 2019 expands to include new digital as well as traditional paper formats. The bridge data will also allow for reexamining the results of the Item Equivalence Study based on nationally representative student populations, as well as addressing potential issues of item model-data misfit and bias in the digital trend items.

By contributing their bridge data to the linking process described in Fig. 4, each eTIMSS country adds to the stability of the eTIMSS-paperTIMSS link, which is based on having equivalent student samples taking items in both eTIMSS and paperTIMSS formats. The linear transformation that establishes this link and adjusts for the mode effect is a global transformation applied in the same way for each country. It makes no provision for differential country-by-country mode effects. However, there is some evidence from the Item Equivalence Study that the mode effect may be stronger in some countries than others, and the bridge data provides each country with an avenue for exploring this issue. By comparing the performance of its students on the eTIMSS and paperTIMSS versions of the trend items, each country can develop a detailed picture of how the mode effect may be operating among its students and which items, if any, are contributing to this effect.

Abbreviations

TIMSS: Trends in International Mathematics and Science Study; IEA: International Association for the Evaluation of Educational Achievement; PSIs: Problem-Solving and Inquiry Tasks.

Authors' contributions

BF, MM, IM, PF contributed to developing the research design, conducting the analysis, and writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors are indebted to psychometric staff at Educational Testing Service (ETS) —Scott Davis, Jonathan Weeks, Ed Kulick, John Mazzeo and Tim Davey—for their assistance with this project. In particular, ETS staff advised on the study design and analytic approach, conducted a range of item-by-item analyses, and implemented the IRT achievement scaling.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Not applicable.

Funding
Not applicable.

Publisher's Note

Received: 13 July 2018 Accepted: 17 October 2018
Published online: 29 October 2018

References

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: American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessment (APA).

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.

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.

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.

Fishbein, B. (2018). *Preserving 20 years of TIMSS trend measurements: Early stages in the transition to the eTIMSS assessment* (Doctoral dissertation). Boston College.

Foy, P. (2017). *TIMSS 2015 user guide for the international database.* Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timss.bc.edu/timss2015/international-database/.

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: 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 procedures 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., & 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.

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., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: how big is the 'mode effect' and what has been done about it? *Oxford Review of Education*. https://doi.org/10.1080/03054985.2018.1430025.

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.

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.

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: TIMSS & PIRLS International Study Center, Boston College.

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

Martin, M. O., Mullis, I. V. S., Foy, P. & Hooper, M. (Eds.). (2016b). 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.

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. (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: Chapman & Hall, CRC Press.

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., 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: 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). Switzerland: Springer International Publishing.

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

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics, 51*(9), 1352–1375.

Parshall, C. G., & Kromrey, J. D. (1993). *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.

Pisacreta, D. (2013). *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.

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.

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

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). *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/.

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.

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.