Validity evidence and measurement properties in technology-enhanced items

Part B: Studies

Authors

Saskia Wools
Paul Drijvers
Remco Feskens
Dylan Molenaar
Emmelien van der Scheer

Contents

1. Overview

1.1 This report

In this report seven different validation studies are presented. The validation studies aim to gather validity evidence for the interpretation and use argument that is presented in Part A of the report. The following research questions are addressed in each of the studies:

Study 1: What definition of problem solving guides and underpins the PSI tasks, and what are the foundations of this definition? Is this definition in line with the main views expressed in mathematics education research literature? (Chapter 2)

Study 2: How do the PSI and eTIMSS items relate to the TIMSS 2019 assessment framework and to what extent are PSI items appropriate to measure problem solving? (Chapter 3)

Study 3: What are the psychometric measurement properties of the eTIMSS and PSI test administrations? And do these properties differ? (Chapter 4)

Study 4: To what extent are the measurement properties of the items included in the PSI and eTIMSS administrations related to each other? (Chapter 5)

Study 5: Are the technology-enhanced items psychometrically comparable to the more traditional items and to what extent are these items subject to differential item functioning? (Chapter 6)

Study 6: Does eTIMSS and PSI process data have the potential to reveal difference in response strategies? Can response time explain differences in item performance across countries? (Chapter 7)

Study 7: To what extent can differences in response strategies be used in order to provide information about the weaknesses and strengths of an educational system? (Chapter 8)

Before these research questions are addressed, the context of the studies is discussed in Section 1.2. The conclusion and discussion of the studies are described in part A of the report.

1.2 Context of the studies

The ILSA that will be validated in this study is the 2019 mathematics cycle of TIMSS. The validation efforts are limited to the mathematics part of TIMSS and only data and items for grade 8 students are considered. In the remainder of this chapter, we first describe the TIMSS study in more detail. The studies are part of a study focusing on performing a validation study of the mathematics part of TIMSS.

TIMSS has been used since 1995 to monitor international trends in mathematics and science. In a recurring cycle of four years, fourth and eighth grade students from all over the world are assessed on their mathematical and science proficiency. TIMSS aims to measure what is taught in schools and the context of learning. To this end, the content of the study is developed collaboratively with the participating countries. As part of this process, the curricula of participating countries are analyzed and developed into an assessment framework and accompanying test materials (Hastedt and Rocher, 2020).

The results of TIMSS - mathematics - are presented in terms of relative performance of students by an overall mathematics score and scale scores for both content (Number, Algebra, Geometry, Data and probability) and cognitive domains (Knowing, Applying and Reasoning) (Martin et al., 2017). The scale scores are constructed by using Item Response Theory (IRT). The average achievement scores provide data users with information about how achievement compares among countries and whether scores are improving or declining over time (Martin et al., 2020).

The TIMSS 2019 cycle was the first in which participating countries could choose between two delivery modes: the paper-based paperTIMSS or the computer-based eTIMSS (Mullis et al., 2020). The paper-based items were converted to digital items, while keeping the items as similar as possible. This resulted in item types as drag-and-drop and drop-down menus, but also in items in which a digital line had to be drawn instead of a line by means of a pen or pencil. The division of countries over paperTIMSS and eTIMSS was nearly 50/50, with a few more countries administering eTIMSS (Mullis et al., 2020).

With the possibility to take the digital version of TIMSS 2019, it was decided that for research and innovation purposes an innovative section would be added to the assessment. The aim was to measure problem solving and inquiry (PSI) in a more detailed way through leveraging the possibilities that digital assessments offer. This resulted in the addition of two booklets (Martin et al., 2017) which are referred to as the PSI booklets. As some of these items are technology-enhanced, these were not administered on paper.

The items were divided by means of a matrix sampling approach resulting in a linked design over 14 booklets. An overview of the design of eTIMSS can be found in Figure 1.1.

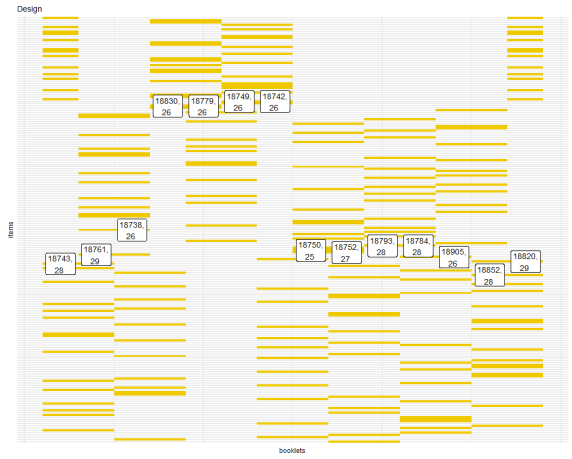


Figure 1.1: Data collection design for mathematics.

Figure 1.1 displays the number of students who were presented with a specific booklet and the number of items that were included in that booklet. For example 18,743 students took booklet 1 and 28 items were included in this booklet. Each item appeared in two booklets.

In the TIMSS 2019 cycle, items were added that should enhance the coverage of problem solving and inquiry (PSI) processes. The items were designed on the basis of the same assessment framework but additional efforts were made. For mathematics, grade 8, this resulted in three PSI tasks divided over two booklets.

Each task started with a description of a problem. Subsequently, items were presented that were linked or related to the main problem. The grade 8 tasks were: Building, Robots and Dinosaur Speed (secure). Building and Robots were presented together in one booklet, Dinosaur speed was the only PSI context in a booklet of its own. More detailed information about these items can be found in the study 'Findings from the TIMSS 2019 Problem Solving and Inquiry Tasks' (Mullis et al., 2021).

The design of the PSI tasks can be found in Figure 1.2 and is presented in a similar way as the design for eTIMSS (Figure 1.1). The two booklets contained the same three tasks presented in a different order. Each booklet contained approximately 30 items and was administered by approximately 10,000 students. Important to note is that the eTIMSS and PSI were *not* connected using a common item design, as will be further discussed in Chapter 5[1].

[1]The design is, however, connected by assuming equivalent samples. This assumption is plausible given the spiraling of the booklets within classrooms.

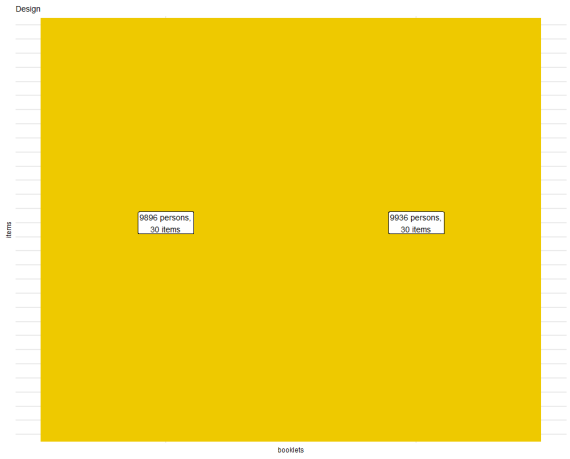


Figure 1.2: Data collection design for PSI.

2. Study 1 – Defining problem solving

2.1 Aim Study 1

To assess the validity of the TIMSS PSI test, the first step is to define how problem solving is understood. What definition of problem solving guides and underpins the PSI test, and what are the foundations of this definition? Is this definition in line with the main views expressed in mathematics education research literature? The aim of Study 1 was to answer these questions.

2.2 Method Study 1

To answer the Study 1 questions, we carried out a document analysis, including TIMSS documentation on the interpretation of problem solving which served as a guideline for the test development. We then compared the results from this analysis with the results of a limited literature review on problem solving in mathematics education research literature.

2.3 Results Study 1

The TIMSS documentation provides relevant information about the definition and understanding of problem solving, that underpins the test design. In Mullis et al. (2021) this is summarized, and we find some criteria for an item to be qualified as problem solving:

> Each PSI task should be situated in a real world, problem, investigation, or activity that provides an underlying narrative or theme for the items. The problem or situation must be sufficiently wide to encompass a number of content and cognitive areas in the Mathematics or Science Frameworks. As much as possible, PSI tasks should attempt to include items addressing various content topics and a range of cognitive demands.[...] The narrative should

> provide a logical or chronological progression from the first item to the ending. [. . .] Because PSI tasks with a single narrative from start to finish can be hard to achieve, PSI tasks also can be written that do not have much narrative, provided there is a common theme to link the items together. The thematic type of PSI task gives students an opportunity to interact with various aspects of a scenario without the order of the interactions having an impact. The items can be independent, while still being coherent and engaging. (p. 5)

Further on, Mullis et al. (2021) phrase:

> Re-imagined for TIMSS 2019, PSI tasks are visually attractive, interactive scenarios that present students with adaptive and responsive ways to follow a series of steps (assessment items) toward a solution or goal. The students' responses are provided via a mixture of selection and constructed response items as well as through various innovative formats to capture students' responses (e.g., number pad, drag and drop, graphing tools, and free drawings). (p. 5)

A third and final quote from the TIMSS documentation highlights the focus on applying and reasoning in PSI: "The tasks covered a range of mathematics and science content domain topics and, consistent with the goal of the PSIs to assess higher-order skills, the majority of the items in the PSIs involved applying and reasoning." (Cotter et al., 2020, p.130).

To summarize, the TIMSS view on problem-solving items is that items should be situated in real life and provide a theme, a common link, or a chronologically progressing narrative. They should cover a range of cognitive domains, focus on higher-order skills such as applying and reasoning, and include a series of steps toward a solution or goal. Let us label this view the narrative view.

How does this view relate to what can be found in mathematics education research literature, in which problem solving is a widely addressed topic? Summarizing the literature in a systematic way is beyond the scope of this report. We can, though, distill some main characteristics of problem-solving tasks that are widely shared, according to some main sources:

- *Non-routine task*: To invite problem-solving activity, the problem at stake should be one that is a non-routine task for the students, that in some sense is new to them, and to which they don't have a standard solving procedure in their toolbox yet (Santos-Trigo, 2020; Schoenfeld, 2015; Selden et al., 1989). This is the most important characteristic of a problem-solving activity, which is of course relative to the students' preliminary knowledge. As curricula may differ in different countries, there is a possibility that a TIMSS item appeals for problem solving in one country and doesn't in another. Still, such cases are expected to be rare.
- *Meaningful situation*: The problem situations should be meaningful to the students. This does not necessarily imply that the problems involve applications or real-life situations. For example, Polyá (2004) in his seminal work mainly used pure mathematical problem situations. There is, however, a tendency to address the modelling of applied situations in problem solving (Doorman et al., 2007).
- *Multi-step and multi-faceted solving process*: Problem solving is considered a complex process that involves orientation to the problem, making a (usually multi-step) plan,

> carrying out the plan, coordinating the different steps that together solve the central problem, monitoring the progress and the overall proceedings, and evaluating and reflecting on the result. Often, a problem-solving task allows for more than one approach, may lead to different processes and even different solutions, and may ask for connecting different mathematical domains and combining different techniques (e.g., applying an algebraic method to a geometric situation).

In short, mathematics education research literature highlights that problem solving involves encountering meaningful non-routine problem situations, which can be tackled through multi-faceted and multi-step solution processes. We label this view on problem solving the non-routine view, and consider it the common view in the mathematics education community. This view is reflected in the PISA definition: "Problem-solving competency is defined as an individual's capacity to engage in cognitive processing to understand and resolve problem situations where a method of solution is not immediately obvious" (OECD, 2017, p.134).

How does this non-routine view on problem solving compare to the TIMSS narrative perspective? The two views are quite different. The TIMSS description does not highlight the aspects of non-routine, meaningful, multi-faceted and multi-step tasks identified above. Rather, it stresses the notion of a "sufficiently wide situation", preferably from real life, that offers a theme or a narrative which serves as an "umbrella" to address different content topics and cognitive demands. Key to a problem-solving task, according to TIMSS, is that there is a theme that may act as a means to cluster, to "glue together" different items, each focusing on different topics. Also remarkable is that the above definition does not stress a main overarching question to be answered, but highlights a theme or a narrative. This narrative view on problem solving is clearly quite different from the non-routine view: as our analysis in Study 2 will show, items that fit the one view do not necessarily fit the other, and vice versa.

Assessing complex skills such as problem solving is a challenge to test designers. The non-routine and multi-step character of problem-solving tasks, particularly in the non-routine view, may be at odds with other test criteria, such as the independency of items (respondents should not suffer from failure on one item in answering the next one), the avoidance of riders, the desired one-step character of items (in multi-step items, it is hard to assess what went wrong), and the suitability for automated scoring. It is these factors, we conjecture, that might have caused the TIMSS narrative view on problem solving. From a validity perspective, however, one might regret that the TIMSS interpretation of problem solving differs from what is common in the mathematics education community. This clearly challenges the validity of TIMSS PSI for the non-routine interpretation of problem solving. This is why Study 2 will integrate both interpretations of problem solving in its analysis.

As a closing activity within Study 1, we also investigated how problem-solving elements could be traced in eTIMSS items. It is interesting to notice that problem solving in the non-routine interpretation does appear in some of the eTIMSS items. For example, consider item M52146 in which the number of matches for a figure has to be determined, based on other sequential figures. Our expectation is that most TIMSS candidates are not very familiar with this type of pattern recognition, and that this is a non-routine task to many of them. Even if the problem situation is not a real-world problem in the sense of encountering

this every day, it clearly makes sense and is experimentally real (Gravemeijer and Doorman, 1999). Also, the solution procedure involves several steps, and may follow different pathways, ranging from numerical attempts to inductive reasoning. As such, this eTIMSS task appeals to students' problem-solving skills in the non-routine view. However, this is not the narrative view that guided PSI, in which we, therefore, don't find this type of items.

3. Study 2 - Qualitative item analysis

3.1 Aim Study 2

PSI tasks are developed to help to "enhance and extend the breadth of the TIMSS assessment to provide more comprehensive coverage of problem solving and inquiry as already described in the assessment frameworks" (Mullis et al., 2021, page 1). The aim of this qualitative study is to investigate how the PSI and eTIMSS items relate to the TIMSS 2019 assessment framework and the extent to which the PSI items are appropriate to measure problem solving.

3.2 Method Study 2

To investigate how the PSI and eTIMSS items relate to the TIMSS 2019 assessment framework, we assigned each mathematics grade 8 eTIMSS and PSI item to the content and cognitive domains of the framework. The following documents were consulted to provide insight in this:

- The TIMSS 2019 Assessment Framework, Chapter 1 - TIMSS 2019 Mathematics Framework (Lindquist et al., 2017).
- TIMSS 2019 Technical Report, Chapter 1 – Developing the TIMSS 2019 Mathematics and Science Achievement Instruments (Cotter et al., 2020).
- The TIMSS 2019 Item Information – Grade 8 the item characteristics (Fishbein et al., 2021).

To zoom in on the PSI items, a qualitative analysis was carried out for each of the three cluster items Building, Robot and T-Rex. The items were analyzed on different aspects:

- Match with the TIMSS view on problem solving.
- Match with the non-routine view on problem solving. An item was marked as requiring problem solving skills according to the non-routine definition when it met

the following conditions: non-routine task, meaningful, multi-step and multi-faceted.
- Match with the "avoid rider" criterion, i.e., the extent to which the consecutive items are independent from each other (Mullis et al., 2021). This was done by evaluating whether an item from a previous item was required for the next (a rider).
- Check whether consecutive items included contradictory information.

The results of the qualitative analysis are summarized in a table for each cluster item, and explained in more detail in the running text.

3.3 Results Study 2

3.3.1 Division eTIMSS and PSI items over content and cognitive domains

eTIMSS and PSI items grade 8 are divided into four content domains: Geometry, Algebra, Number, and Data and probability. The division of eTIMSS and PSI items for mathematics grade 8 is based on the item information, instead of the reported information in the Technical Report. In this report a design-based inference was used, instead of a model-based inference (Robitzsch and Lüdtke, 2022). Whereas a design-based inference is based on a sampling design for persons and items, a model-based inference is based on specific assumptions of statistical models. Robitzsch and Lüdtke argue that a design-based approach is preferred for the reporting of ILSA results.

Table 3.1: Overview of number of items per content domain of the eTIMSS and PSI items for mathematics grade 8.

Content domain	**eTIMSS**			**PSI**	
	N	Target %*	%**	N	%
Algebra	55	30	29.1	14	46.7
Data and probability	33	20	17.5	4	13.3
Geometry	41	20	21.7	11	36.7
Number	60	30	31.8	1	3.3

** Lindquist, Philpot, Mullis & Cotter (2017)*

*** These numbers deviate from the realized numbers as presented in Cotter, Centurino & Mullis (2020) due to taking of the stem of the item as the unit of analysis.*

Table 3.1 shows that all content domains are represented in both the eTIMSS and PSI items, and that for eTIMSS the targeted percentages are reached. When eTIMSS and PSI items are compared there is a clear difference in the division of items over the content domains.

With regard to the cognitive domains, items were divided into the cognitive domains Knowing, Applying and Reasoning. Table 3.2 presents how eTIMSS items and PSI items are divided into these cognitive domains, similar to the content domains.

Table 3.2: Overview of number of items per cognitive domain of the eTIMSS and PSI items for mathematics grade 8.

Cognitive domain	eTIMSS			PSI	
	N	Target %*	%**	N	%
Applying	87	40	46.0	18	60
Knowing	62	35	32.8	0	0
Reasoning	40	25	21.1	12	40

** Lindquist, Philpot, Mullis & Cotter (2017)*

*** These numbers deviate from the realized numbers as presented in Cotter, Centurino & Mullis (2020) due to taking of the stem of the item as the unit of analysis.*

As shown in Table 3.2 the realized percentage of items for eTIMSS is in line with the target percentage. The cognitive domains for the PSI items are Applying and Reasoning, while none of the items relate to the cognitive domain Knowing. This is, according to Cotter, Centurino & Mullis (2020) in line with the goal of PSIs to assess higher-order skills.

3.3.2 Qualitative cluster item analysis

The Building cluster item

The Building cluster item is described as follows: "Students determine the dimensions of a shed, to store equipment, including a barrel to collect rainwater" (Cotter et al., 2020 p. 1.31). It is considered "a mathematics problem, where students work from a visualization to a finished product involving multiple steps and evaluation of interim results" (p. 5). Table 3.3 provides an overview of the analysis results.

Table 3.3: Overview of the qualitative analysis results for the Building cluster item.

PSI item	Description	PS* TIMSS	PS* non routine	Rider	Contradictory information
MQ12B01	Building Size				
MQ12B02	Roof - expression				
MQ12B03	Construction the walls	x			
MQ12B04A	Painting the walls – area of side wall	x			
MQ12B04B	Painting the walls – total area	x		x	
MQ12B04C	Painting the walls – cost of paint	x	x		x
MQ12B05A	Water tank – volume of tank				
MQ12B05B	Water tank – greater volume of tank				
MQ12B05C	Water tank –volume of tank cylinder		x		

** PS = problem solving*

MQ12B01 (Screen 2) concerns the floor area of the building, and as such does not contribute to the overall narrative of deciding how much paint is needed to paint the walls. Also if the three units were merged, the middle supporting poles would be twice as thick as shown in the other screens.

MQ12B02 concerns the roof. Again, this is not contributing to solving the central paint question. Also, it seems artificial to express the length of a roof in terms of exact value containing a square root.

In MQ12B03 the narrative is more functional as it focuses on the walls. Also, the task allows for different solutions. Still, it is relatively routine, and not problem solving in the non-routine interpretation.

MQ12B04A fits well in the narrative and as such is considered problem solving in the TIMSS interpretation. MQ12B04B involves a rider: if students make a mistake in 4A, then

they might multiply this incorrect value by 2 and this would lead to an incorrect answer for item 4B. This might, however, be taken into account through intelligent automated scoring.

MQ12B04C raises another issue. Even if it is considered problem solving in both interpretations, students might be hindered by the total area of 120 m2 to be painted, whereas they calculated in 4B that the total area is 88 m2. This inconsistency might worry students, who may doubt whether their answer to 4B was correct.

The water tank items MQ12B05 A-B-C are not really related to the narrative of building a store building, and as such do not exemplify the TIMSS view on problem solving very well. Item 5C, however, does involve non-routine reasoning with volume as a function of radius, which is likely to be a non-routine task for these students.

To summarize, the narrative in the Building cluster only weakly connects the different items, and as such doesn't really add to the separate items. The problem situation is not a real-live context to most of the students. The issues with the rider and the inconsistency may threaten the item's validity.

The Robot cluster item

The Robot cluster item is described as follows: "Students determine the functions using a robot that applies a function to determine y for any given value of x" (Cotter et al., 2020, p. 1.31). Table 3.4 provides an overview of the analysis results.

Table 3.4: Overview of the qualitative analysis results for the Robot cluster item.

PSI item	Description	PS TIMSS	PS non routine	Rider	Contradictory information
MQ12R01	Robots - rule		x		
MQ12R02A	Robots – table for X				
MQ12R02B	Robots – table for Y			Possibly	
MQ12R02C	Robots – different rule		x	x	

MQ12R01 concerns a "guess my rule" task, that is not uncommon in early algebra education, but probably not routine to the TIMSS candidates. Therefore, it is considered problem solving in the non-routine sense. From a narrative perspective, the robot does not add much to the task, and as such is a somewhat artificial additional element.

MQ12R02 concerns similar issues, but might have a small rider effect, as the correct answers to the table tasks (2A and 2B) are very helpful in finding the answer to 2C.

To summarize, the narrative in the Robot cluster is weak and artificial. In fact, the items could very well be part of the eTIMSS tests. From the non-routine perspective, they do invite non-routine problem solving. It is unclear how they address the TIMSS interpretation of problem solving.

The T-Rex cluster item

The T-Rex cluster item is described as follows: "Students use the relationships between foot length, leg height, and stride length to estimate how fast a dinosaur could run" (Cotter et al., 2020, p. 1.31). Table 3.5 provides an overview of the analysis results. A detailed description for each item cannot be provided in this report as the item content is restricted.

Table 3.5: Overview of the qualitative analysis results for the T-Rex cluster item.

PSI item	Description	PS TIMSS	PS non routine	Rider	Contradictory information
MQ12D01	Footprint length to foot length				
MQ12D02AA	Relationship between foot length and leg height - 37				
MQ12D02AB	Relationship between foot length and leg height - 61				
MQ12D02AC	Relationship between foot length and leg height - 78				
MQ12D02B	Relationship between foot length and leg height - formula	x	x		
MQ12D03	Trexy's leg height	x	x		x
MQ12D04A	Identifying left and right footprints – bottom angle				
MQ12D04B	Identifying left and right footprints – foot				
MQ12D05	Trexy's stride length				
MQ12D06A	Ratio for a dinosaur's movement – shortest stride length	x	x		
MQ12D06B	Ratio for a dinosaur's movement – longest stride length	x	x		
MQ12D07A	Calculate Trexy's running speed – running speed				x
MQ12D07B	Calculate Trexy's running speed – faster speed	x	x		
MQ12D08A	How fast could trexy run a 100m race? - 40 meters	x	x		
MQ12D08B	How fast could trexy run a 100m race? - entire race	x	x		

To summarize, the narrative in the T-Rex cluster is more or less clear, with the running speed as the main goal. There are clearly some activities that can be classified as problem solving, both in the TIMSS interpretation and in the non-routine view. Still, the task and its narrative altogether are quite long, and some elements don't really contribute to the main goal (e.g., the items MQ12D04 and MQ12D05), whereas others are somewhat confusing in terms of consistency (items MQ12D03 and MQ12D07A).

As an overall conclusion, the qualitative analysis of the PSI items shows:

- There is a match with the TIMSS view on problem solving. This match might be stronger for the Building and the T-Rex cluster, and is not very present in the Robot cluster.
- There is a match with the non-routine view on problem solving. Again, this should be further elaborated, and in some items it is missing.
- There are a limited number of riders.
- Some consecutive items include contradictory information.

4. Study 3 - Psychometric analysis

4.1 Aim Study 3

In this study we conducted standard psychometric checks on the eTIMSS and PSI items. We specifically studied whether the reliability of the eTIMSS and PSI scales are sufficient to enable inferences about cross-country differences, and if the unidimensionality of the eTIMSS and PSI scales (as an aspect of validity) are strong enough (see Paragraph 4.3.2) to represent the content domain of mathematics. We were especially interested in differences between the eTIMSS and PSI scales. That is, as the PSI items are explicitly developed as digitally based interactive assessment items, it is relevant to see if this interactive assessment mode of assessment affects the psychometric properties of the scale. Finally, we compared the two scales on the nonresponse as an indicator of motivation.

4.2 Method Study 3

4.2.1 Reliability of the eTIMSS and PSI scales

We considered the reliability of the PSI items in terms of classical test theory statistics (Cronbach's alpha and corrected item-total correlations). To enable an interpretation of the results, we contrasted the results from the PSI scale to those of the eTIMSS booklet 1 items from the same countries that administered the PSI items. As the PSI item data was collected using an equivalent groups design (see Martin et al., 2020), the results can be validly compared to arrive at a conclusion concerning the differences between the (technology-enhanced) PSI items, and the (more common) eTIMSS items. "Partially correct"-responses are scored as incorrect. For the reliability analysis, all items were treated separately (i.e., items that are interdependent are treated as separate items in this section). As a result, the PSI scale contained 28 binary items, and the eTIMSS subset of items contained 31 binary items.

4.2.2 Unidimensionality of the eTIMSS and PSI scales

To study the structure of the item scores, a unidimensional two-parameter item factor model was fit to the data from each country separately. To account for interdependencies between items (i.e., items that are in the same screen), these items were summed and represented as a single item in the model. As a result, the PSI scale contained 15 items, and the eTIMSS scale contained 28 items. The fit of the unidimensional model, and thus the tenability of unidimensionality, was assessed by the Root Mean Square Error of Approximation (RMSEA, Browne and Cudeck, 1992), the Comparative Fit Index (CFI; Bentler and Bonett, 1980) and the Tucker-Lewis Index (TLI; Bentler and Bonett, 1980). We used the guidelines by Schermelleh-Engel, Moosbrugger, & Müller (2003), which means that we consider the RMSEA to be 'acceptable' for values between 0.08 and 0.05, and 'good' for values smaller than 0.05. For the CFI and TLI we considered values between 0.95 and 0.97 as 'acceptable' and values above 0.97 as 'good'.

4.2.3 Analysis of nonresponse

To provide an overview of the nonresponse the findings as presented in Mullis et al. (2021) are described.

4.3 Results Study 3

4.3.1 Reliability of the eTIMSS and PSI scales

See Table 4.1 for the reliability estimates by means of Cronbach's alpha and the mean corrected item-total correlation. In addition, see Figure 4.1 for a graphical representation of these results. As can be seen from the table and figure, the estimated reliability of the PSI items are comparable to the eTIMSS items in magnitude, indicating that both scales are hardly different in reliability.

In Figure 4.1 a graphical display of the differences in Cronbach's alpha and mean corrected item-total correlations (Mean r) of the PSI scale and the eTIMSS scale across countries can be found. The green lines denote the PSI scale, blue lines denote the eTIMSS scale. The country numbers correspond to the numbering used in Table 4.1.

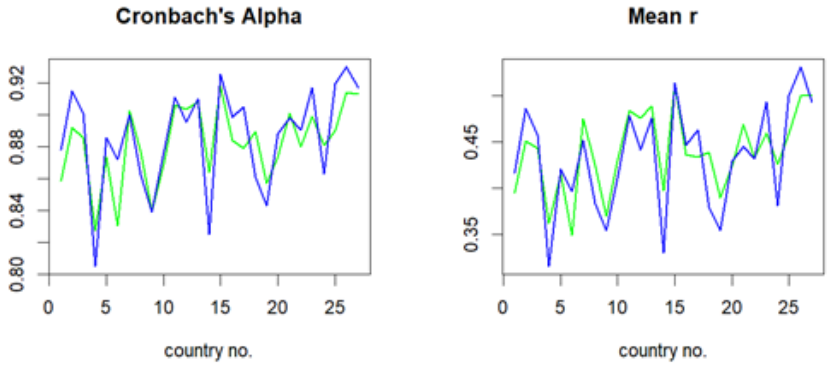


Figure 4.1: Cronbach's alpha and mean corrected item-total correlations across scales and countries.

Table 4.1: Reliability estimates of the eTIMSS and the PSI scale across countries.

		eTIMSS		**PSI**	
No.	*Name*	*Alpha*	*Mean r'*	*Alpha*	*Mean r'*
1	AAD	0.88	0.42	0.86	0.39
2	ADU	0.92	0.49	0.89	0.45
3	ARE	0.90	0.46	0.89	0.44
4	CHL	0.80	0.31	0.83	0.36
5	COT	0.89	0.42	0.87	0.41
6	CQU	0.87	0.40	0.83	0.35
7	ENG	0.90	0.45	0.90	0.47
8	FIN	0.86	0.38	0.88	0.42
9	FRA	0.84	0.35	0.84	0.37
10	GEO	0.88	0.41	0.87	0.43
11	HKG	0.91	0.48	0.91	0.48
12	HUN	0.90	0.44	0.90	0.48
13	ISR	0.91	0.48	0.91	0.49
14	ITA	0.82	0.33	0.86	0.40
15	KOR	0.93	0.51	0.92	0.51
16	LTU	0.90	0.45	0.88	0.44
17	MYS	0.90	0.46	0.88	0.43
18	NOR	0.86	0.38	0.89	0.44
19	PRT	0.84	0.35	0.86	0.39
20	QAT	0.89	0.43	0.87	0.42
21	RMO	0.90	0.44	0.90	0.47
22	RUS	0.89	0.43	0.88	0.43
23	SGP	0.92	0.49	0.90	0.46
24	SWE	0.86	0.38	0.88	0.43
25	TUR	0.92	0.50	0.89	0.46
26	TWN	0.93	0.53	0.91	0.50
27	USA	0.92	0.49	0.91	0.50

4.3.2 Unidimensionality of the eTIMSS and the PSI scales

Table 4.2 contains the RMSEA, CFI, and TLI of the eTIMSS and the PSI scales across countries. In Figure 4.2, the results are graphically depicted. Unidimensionality is generally tenable for both scales. That is, the TLI/CFI is above 0.97 and the RMSEA is below 0.05 in all countries. Statistically, both scales can be considered unidimensional. However, an interesting difference -which is best notable from the RMSEA - is that the eTIMSS seems to fit better to a unidimensional model than PSI.

Table 4.2: Fit of a unidimensional latent variable model to the PSI and eTIMSS scales for each country.

		PSI			eTIMSS		
No.	*Name*	*CFI*	*TLI*	*RMSEA*	*CFI*	*TLI*	*RMSEA*
1	AAD	0.98	0.98	0.04	1.00	1.00	0.00
2	ADU	0.99	0.99	0.04	1.00	1.00	0.00
3	ARE	0.99	0.98	0.04	1.00	1.00	0.02
4	CHL	0.99	0.99	0.03	1.00	1.02	0.00
5	COT	0.99	0.99	0.03	1.00	1.00	0.00
6	CQU	0.99	0.99	0.03	1.00	1.01	0.00
7	ENG	1.00	1.00	0.00	1.00	1.01	0.00
8	FIN	0.99	0.99	0.03	1.00	1.00	0.00
9	FRA	1.00	1.00	0.01	1.00	1.00	0.00
10	GEO	1.00	1.01	0.00	1.00	1.01	0.00
11	HKG	1.00	1.00	0.02	1.00	1.01	0.00
12	HUN	1.00	1.00	0.01	1.00	1.01	0.00
13	ISR	0.99	0.99	0.04	1.00	1.01	0.00
14	ITA	0.99	0.98	0.04	1.00	1.01	0.00
15	KOR	1.00	0.99	0.04	1.00	1.01	0.00
16	LTU	1.00	1.00	0.01	1.00	1.01	0.00
17	MYS	0.99	0.99	0.03	1.00	1.00	0.00
18	NOR	0.99	0.99	0.03	1.00	1.01	0.00
19	PRT	1.00	1.00	0.02	1.00	1.00	0.00
20	QAT	1.00	1.00	0.01	1.00	1.01	0.00
21	RMO	1.00	1.00	0.03	1.00	1.01	0.00
22	RUS	0.99	0.98	0.04	1.00	1.00	0.01
23	SGP	0.99	0.99	0.04	1.00	1.01	0.00
24	SWE	0.98	0.98	0.05	1.00	1.01	0.00
25	TUR	1.00	1.00	0.02	1.00	1.00	0.01
26	TWN	1.00	1.00	0.02	1.00	1.00	0.00
27	USA	1.00	1.00	0.03	1.00	1.00	0.00

In Figure 4.2 the fit statistics of the one-dimensional latent variable model are presented. Green lines denote the PSI scale, blue lines denote the eTIMSS scale. The country numbers correspond to the numbering used in Table 4.2.

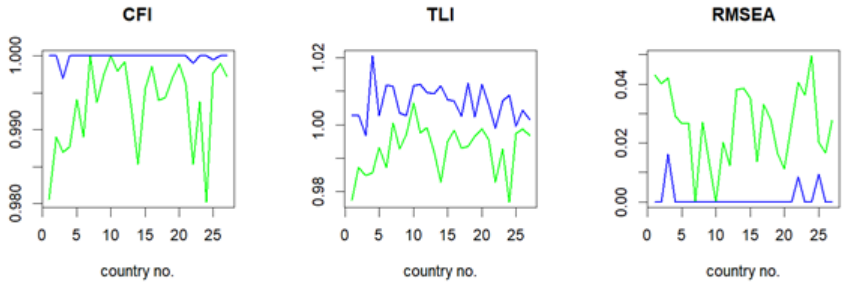


Figure 4.2: Fit statistics across scales and countries.

4.3.3 Is there any significant nonresponse?

In the PSI report it is described extensively that the nonresponse is higher for PSI items than for regular e-TIMSS items (Mullis et al., 2021). Mullis et al. (2021) found that 83 percent of the students reached all PSI items, while 94 percent of the eTIMSS students reached all items. A similar percentage of students experienced a lack of time (respectively 2 percent for eTIMSS and 3 percent for PSI of the students). However, a difference was found for students who stopped responding, as 14 percent of the students who made the PSI tasks stopped responding while only 4 percent of the students who made eTIMSS items stopped responding. Mullis et al. (2021) suggest that this can be explained by a lack of motivation or fatigue, but it was also found that students who stopped responding were associated with lower performance.

5. Study 4 – Comparing PSI & eTIMSS

5.1 Aim Study 4

The three mathematics PSI tasks included for the grade 8 students in TIMSS 2019 should call on students to integrate and apply process skills and content knowledge to solve mathematics problems[1]. Within this study we evaluated to what extent the measurement properties of the items included in the PSI and eTIMSS administrations are related to each other.

5.2 Method Study 4

In order to psychometrically compare the PSI outcomes with the eTIMSS outcomes, we conducted multiple calibrations to equate the test results. Equating is a statistical process in which scores on different tests or booklets are adjusted in order to make these scores comparable (Kolen and Brennan, 2013). In an Item Response Theory model (IRT, Lord and Novick, 2008), the abilities of students and the item characteristics are related to each other in a common probabilistic model using separate parameters. If a particular set of items applies to a particular item response model, the person parameters can be expressed on the same scale based on subsets of the items. The person parameters can then be determined independently from the test, and there is no further need to equate the scores: the scores are already expressed on the existing scale (Engelen and Eggen, 1993). This does not imply that there are no equating issues. First of all, the items should fit a particular item response model. Before equating can take place, the calibration problem has to be solved. Calibration means the choice for a particular IRT model, the collection of data following a particular design, item estimation and testing the validity of the model.

A crucial step in the calibration of the PSI and the eTIMSS items was taking into account

[1] https://www.iea.nl/news-events/news/interactive-timss-2019-pioneers-digital-assessment

that these administrations did not have overlapping anchor items. This implied that we could not make use of conditional maximum likelihood (CML) item parameter estimation, but were forced to use marginal maximum likelihood (MML) estimation and make an additional assumption about the population distribution. The eTIMSS did however have overlapping items and a connected design (see Section 1.2) and for these booklets we could apply a concurrent calibration without needing to make additional assumptions.

For this first step in the calibration we made use of an extended version of the Nominal Response Model (NRM, Bock, 1972; Maris et al., 2015; Cressie and Holland, 1983), where the scoring parameters for the response categories of an item are known integers, and for which the manifest probabilities can be determined in closed form. The NRM is a generalization of the Rasch (Rasch, 1960) and Partial Credit Model (PCM, Masters, 1982) in which every response category in a polytomous item gets its own score. In the general version of the model, the score is a parameter which is estimated (just like item discrimination in the 2PL). This model has been implemented in the R package *dexter* (Maris et al., 2022). After having estimated the item parameters in the eTIMSS administration, we estimated the item parameters of the PSI items on the same scale by combining the item responses of both the eTIMSS and PSI administrations, we read in and fixated the already estimated eTIMSS items, and we estimated the PSI items using MML estimation (Maris et al., 2020). This procedure could also be applied because students within countries had been randomly allocated to either eTIMSS or PSI and were sampled from the same classrooms (Fishbein and Foy, n.d.). With the PSI an eTIMSS items on a common scale, the outcomes can be compared. Within this study we have included all item responses collected in TIMSS (mathematics) within grade 8 .

As a second step in the comparison of eTIMSS and PSI, we compared the correlational structure of the content domains within the two measurements. In order to do so we had to take the measurement uncertainty of the measurements of these domains into account. The observed correlation coefficients are namely underestimates of the 'true' correlations. That is because random measurement error reduces the (observed) correlation. By using plausible values we were able to estimate the relationship between two constructs as if they were measured perfectly reliably and free from random errors that occur in all observed measures. These results are presented in Figure 5.3.

As a final step in the comparison of the PSI and eTIMSS items, the mathematics assessment expert within this team (PD) identified five item pairs which appeal to the same skills and proficiency of students. It was expected by the mathematics assessment expert that these items should elicit more or less the same response behavior. The items can be found in Table 5.1.

Table 5.1: Comparable items in eTIMSS and PSI.

Item pair	**eTIMSS**	**PSI**
1	M52146B	MQ12R01
2	M52146B	MQ12R02C
3	M62027	MQ12R01
4	M62027	MQ12R02C
5	M72094	MQ12D03

We compared the item characteristic curves (ICCs) for these item pairs visually in order to evaluate if these item pairs invoked approximately the same response behavior (cf. Zumbo, 1999). The PSI item is always presented next to the similar eTIMSS item. The ICC visualizes the probability of a correct response given proficiency. In case of similar response behavior one would expect more or less similar, comparable ICCs for the pairs of items. The results can be found in Figures 5.5 and 5.6.

5.3 Results Study 4

In Figure 5.1 the distributions of item difficulties - referred to as *beta* - across the eTIMSS and PSI mode are displayed.

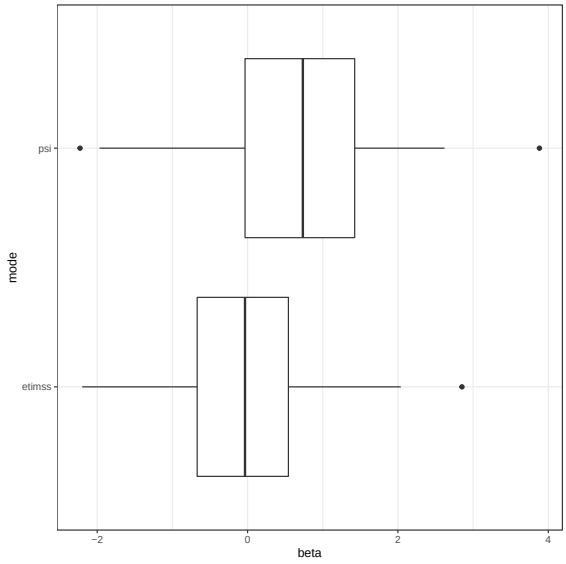


Figure 5.1: Item difficulties in PSI and eTIMSS mode.

As can be seen from Figure 5.1 the previous results are also confirmed in this analysis. It appears that the PSI items are in general substantially more difficult compared to the items included in eTIMSS. In Figure 5.2 we will further evaluate those differences by presenting the item difficulties for both modes categorized by content and cognitive domain.

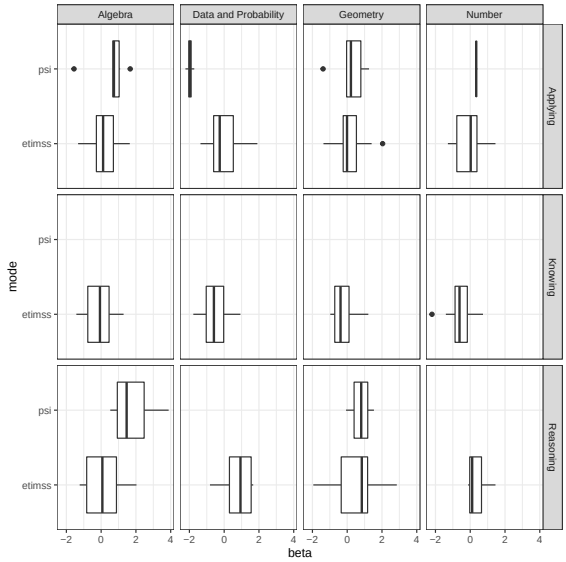
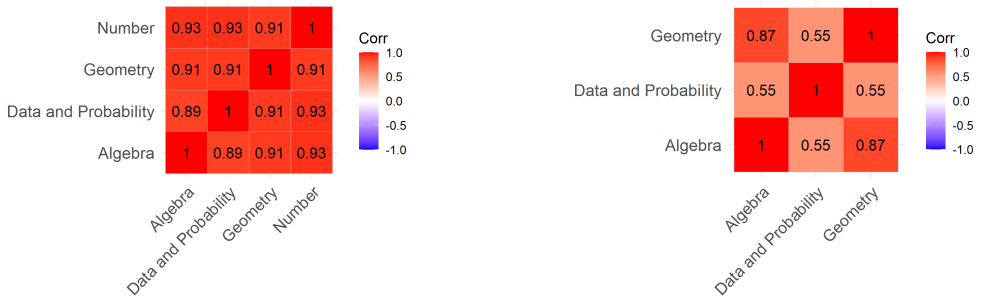


Figure 5.2: Item difficulties of domains in PSI and eTIMSS mode.

In Figure 5.2 the item difficulties across content and cognitive domains are presented. Some of the domains - e.g. the complete "Knowing' domain - are not included in the PSI administration. Probably the largest difference is found in the content domain "Algebra" combined with the cognitive domain "Reasoning". The item difficulties in the PSI administrations are much larger compared to the items measuring the same domains in the eTIMSS domain.

In Figure 5.3 the (latent) correlations coefficients between the content domains within the eTIMSS (Figure 5.3a) and PSI (Figure 5.3b) are displayed.

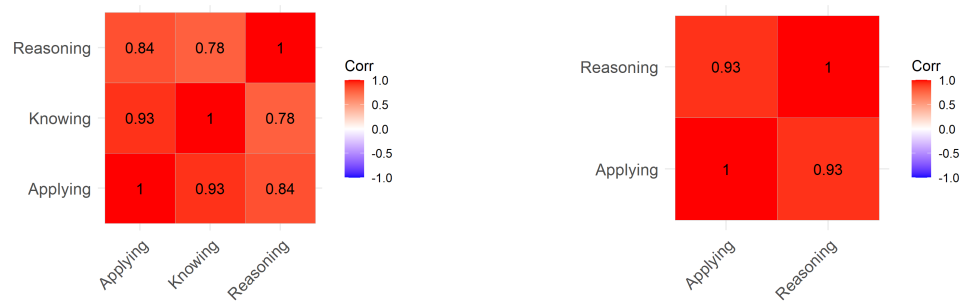


(a) eTIMSS (b) PSI

Figure 5.3: Correlation content domains.

Whereas in the eTIMSS administration all four content domains are included, the PSI administration only contains the domains Algebra, Data and Probability and Geometry. The latent correlations between the domains within eTIMSS are very high - all above 0.9 - suggesting an unidimensional measurement of the construct mathematics (see also Paragraph 4.3.2). Within the PSI tasks is it noticeable that the latent correlation between Data and probability and the two other included domains is much lower (0.55) compared to the other reported correlation coefficients, suggesting that other skills are required to solve these items.

In Figure 5.4 the (latent) correlations coefficients between the cognitive domains within again the eTIMSS (Figure 5.4a) and PSI (Figure 5.4b) can be seen. The latent correlation coefficients between the cognitive domains are reasonably high for the eTIMSS assessment, and high for the PSI assessment. The (latent) correlation coefficient between the cognitive domain Knowing and Reasoning is somewhat lower (0.78) compared to the other reported correlations. The high values of the other coefficients support the claim that these call upon the same ability.



(a) eTIMSS (b) PSI

Figure 5.4: Correlation cognitive domains.

Figures 5.5 and 5.6 display the item characteristic curves (ICCs) for the items which were identified as items which are expected to elicit similar response behavior by the assessment experts. The ICCs of the paired items look relatively comparable. Of course it can be seen from the location of the curves that the items in the PSI administration are more difficult (i.e. one needs "more" proficiency to reach a 0.5 probability to endorse the item) compared to the eTIMSS items, as had already been established previously. A difference that is striking is the higher discrimination ability of the PSI items in the above-average proficiency range compared to the eTIMSS items. There it can be seen that the ICCs tend to be steeper among the PSI items. The eTIMSS items on the other hand show, in general, an increasing ICC over the complete proficiency range. The flatter curves in the lower range of proficiency combined with the steeper patterns (and thus more informative measurements) in the higher range of proficiency could indicate that the PSI items are *less* suited for below-average performing students and, conversely, are *more* suited for above-average performing students.

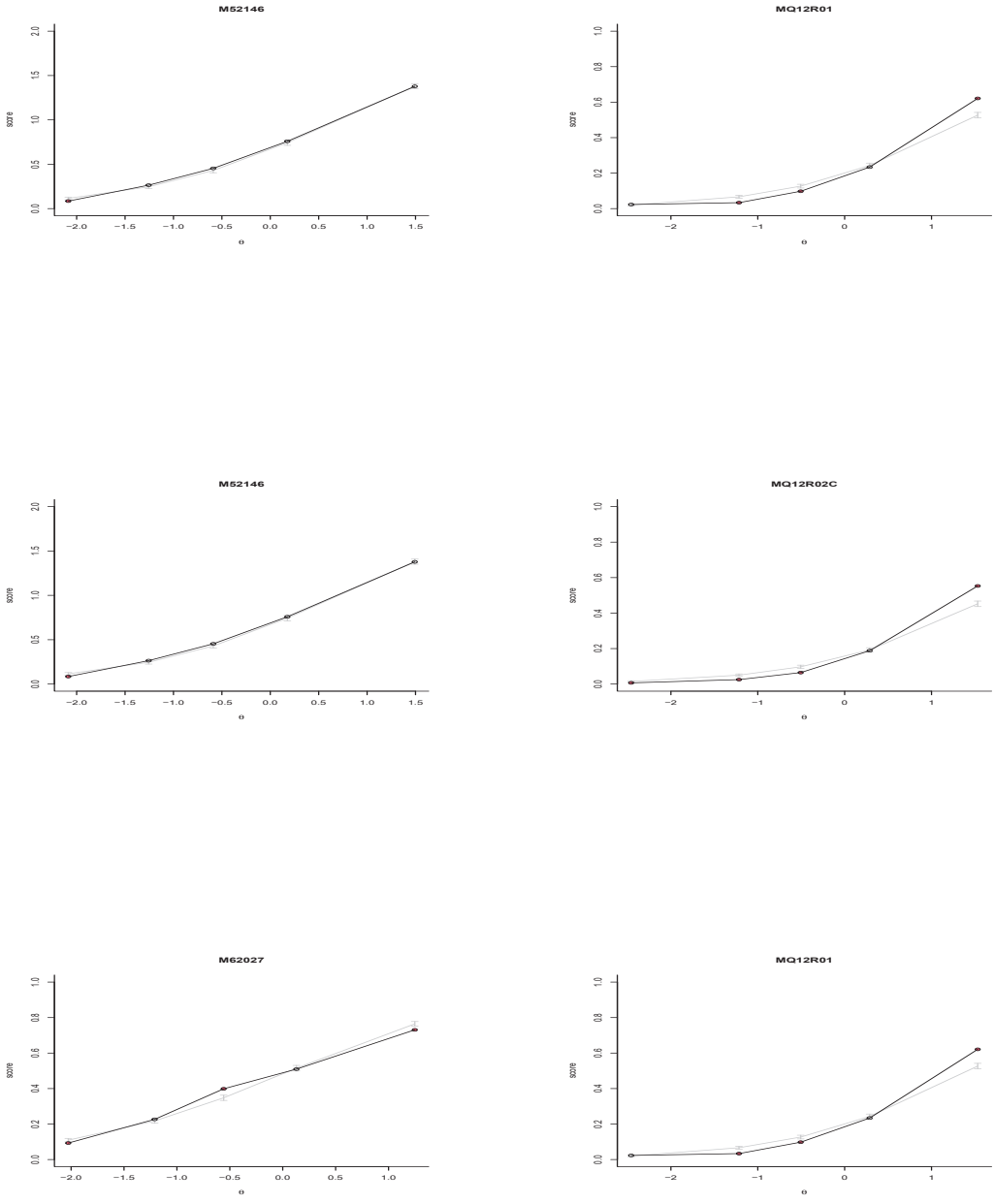


Figure 5.5: Item characteristics curves item pairs 1.

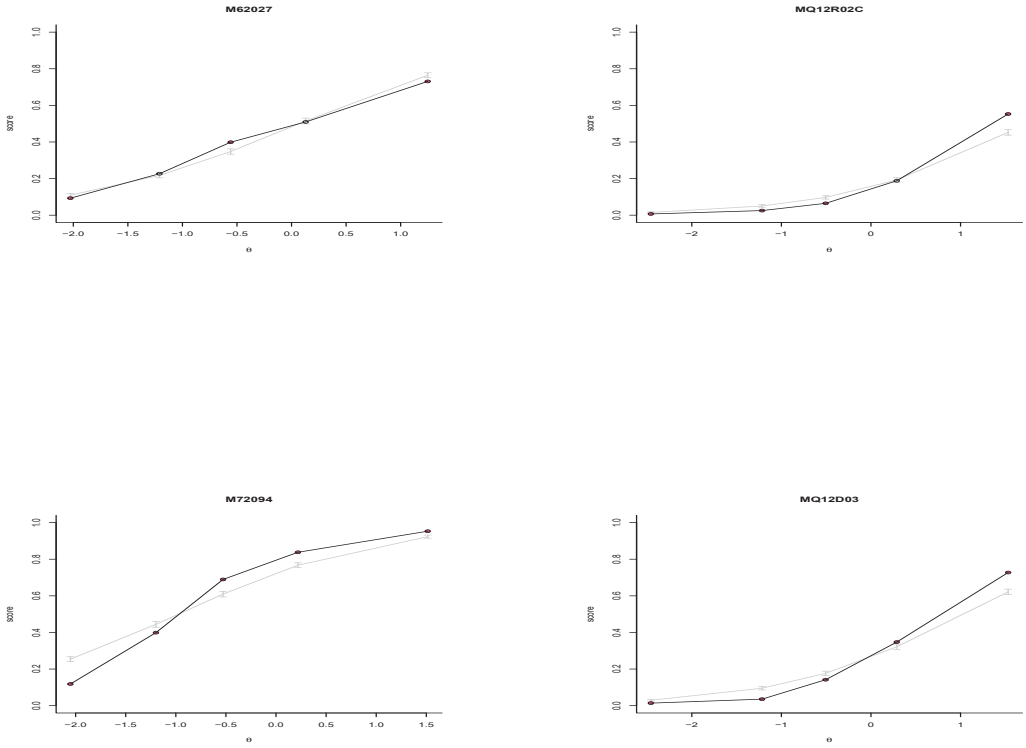


Figure 5.6: Item characteristics curves item pairs 2.

6. Study 5 – Classical differential item functioning

6.1 Aim Study 5

Technology-enhanced items depend on the interactions a respondent has with the computer during the measurement. This might potentially affect the psychometric quality with which mathematics skills are being measured by the items. For example, individual differences in computer skills may confound the mathematics measures or obscure differences in mathematics proficiency. In ILSAs, this may affect comparisons across countries. In addition, the question arises if the technology-enhanced items perform psychometrically comparable to the more traditional items. In this study, we therefore aimed to provide an insight in systematic differences in the difficulty of technology-enhanced items and 'regular' items by means of a differential item functioning (DIF) analysis across countries.

6.2 Method Study 5

As discussed above, if the items of a scale lack DIF, one can meaningfully compare the countries on the underlying proficiency, in this case mathematics. As the requirement of a total absence of differences in the measurement properties across groups (i.e., equal discrimination and equal difficulty parameters across countries) is often argued to be unrealistic if there are many groups or countries (Davidov et al., 2015), we focused on an approximate method to test for DIF, where small differences in the measurement parameters are allowed. Specifically, we used the unidimensional model described above to test for DIF using the so-called alignment method (Muthén and Asparouhov, 2014) across countries in the eTIMSS and PSI item scores[1]. In the alignment method, discrimination parameters and difficulty parameters are linearly transformed to be as close as possible. In the final solution, it can then be seen which parameters are still significantly different

[1]Please note that this model is equivalent to the Graded Response Model.

from each other indicating DIF. In the present project we solely focused on uniform DIF (differences in difficulty across countries), as these effects are best interpretable across many countries.

As for the scales, the sample sizes differ largely across countries. We therefore randomly selected 400 subjects per country to ensure a fair comparison across countries (as larger samples have larger power to detect DIF which would confound our results across countries). In addition, we only focussed on the countries that completed the PSI items, which are 27 countries. To account for multiple testing, in all results of the alignment method we have adopted a level of significance of 0.001. This choice has been based on Muthén and Asparouhov (2014).

6.2.1 Item categorization

The results concerning the DIF analysis were compared across different forms of technology-enhanced items. That is, first, the eTIMSS items and PSI items were categorized according to the following two characteristics:

- *Is technology required to answer the item?* An item was marked as 'required' when an interactive act should be undertaken by a student to answer the item. Thus, for example, drawing a line, drag and drop, sliding a point on a graph or using a ruler.
- *Can technology be used as an aid to answer the item?* An item was marked 'aid' when an interactive act could be undertaken by a student, but when it was not part of the answer. For example, drawing a auxiliary line that could form a step towards a solution.

This categorization scheme resulted in the classification depicted in Table 6.1 for the PSI items and Table 6.2 for the eTIMSS items.

Table 6.1: Overview of PSI items that are technology-enhanced.

Variable	Required	Aid
MQ12B03	x	
MQ12B04A		x
MQ12B04B		x
MQ12R01	x	
MQ12D02AA	x	
MQ12D02AB	x	
MQ12D02AC	x	
MQ12D05	x	
MQ12D08A	x	

In the analysis, a comparison is made between non-technology enhanced PSI items with technology-enhanced PSI items.

Table 6.2: Overview of eTIMSS items that are technology-enhanced.

Variable	Required	Aid	Booklet
M72170	x		1
M62300	x		3
M72181	x		3
M62002	x		5
M52036	x		5
M62288	x		5
M52048	x		7
M72119	x		7
M62296	x		11
M72019	x		11
M72026	x		13
M72103		x	1
M72198A		x	1
M62244		x	3
M62040		x	5
M62287		x	9

To get an insight into whether the measurement of mathematics skill using technology-enhanced items is confounded, the different kinds of technology-enhanced items (i.e., technology as 'aid' or 'required') were compared in terms of 1) the estimated difficulty of the item; and 2) the number of countries in which the item is flagged as DIF item. As shown in Table 6.2, the technology-enhanced items are in some of the eTIMSS booklets. The items from booklets 1, 3, 5 and 7 were selected for the analysis as they entail most of the technology-enhanced items.

6.3 Results Study 5

6.3.1 DIF with respect to difficulty across country

First, in Appendix A.1, the results are depicted with respect to the DIF analysis of the item scores across countries for the eTIMSS scale (see Tables A.1, A.2, A.3, and A.4 for respectively booklet 1, 3, 5, and 7) and for the PSI scale (see Table A.5). Note that the total number of countries equals 27. Some items show DIF in many countries for both the eTIMSS and the PSI scales. However, most importantly, the PSI scale does not seem to show more DIF as compared to the eTIMSS. That is, the average number of DIF countries for eTIMSS equals 6.094 (SD: 3.913) for booklet 1, 5.303 (SD: 4.104) for booklet 3, 3.933 (SD: 3.300) for booklet 5, and 6.219 (SD: 3.590) for booklet 7. For the PSI scale, the average number of DIF countries equals 5.321 (SD: 2.653) which is not substantially different from the eTIMSS scales. Two items from the PSI scale show DIF in relatively many countries (more than 10), items MQ12D02AA and MQ12D03. For the eTIMSS there are respectively 6, 6, 2, and 6 items which show DIF across more than 10 countries for booklets 1, 3, 5, and 7. Of course, the item numbers differ somewhat (respectively 32, 33, 30, 32, for the eTIMSS and 28 for PSI), but most importantly, PSI is at least performing similarly (if not slightly better) than the eTIMSS in terms of the number of DIF items.

See Table A.6 for result with respect to the technology-enhanced items and the more

traditional items. As can be seen for both scales, the technology-enhanced items do not show DIF across countries.

6.3.2 Difficulty of TE items in eTIMSS

Figure 6.1 illustrates what the estimated difficulty of technology-enhanced items is, as compared to the non-technology enhanced items for Booklets 1, 3, 5 and 7 of eTIMSS. The items for which technology was required are shown in orange, items for which technology could be used as aid are shown in in yellow. From Figure 6.1 it can be concluded that no systematic differences regarding the difficulty of the items is found.

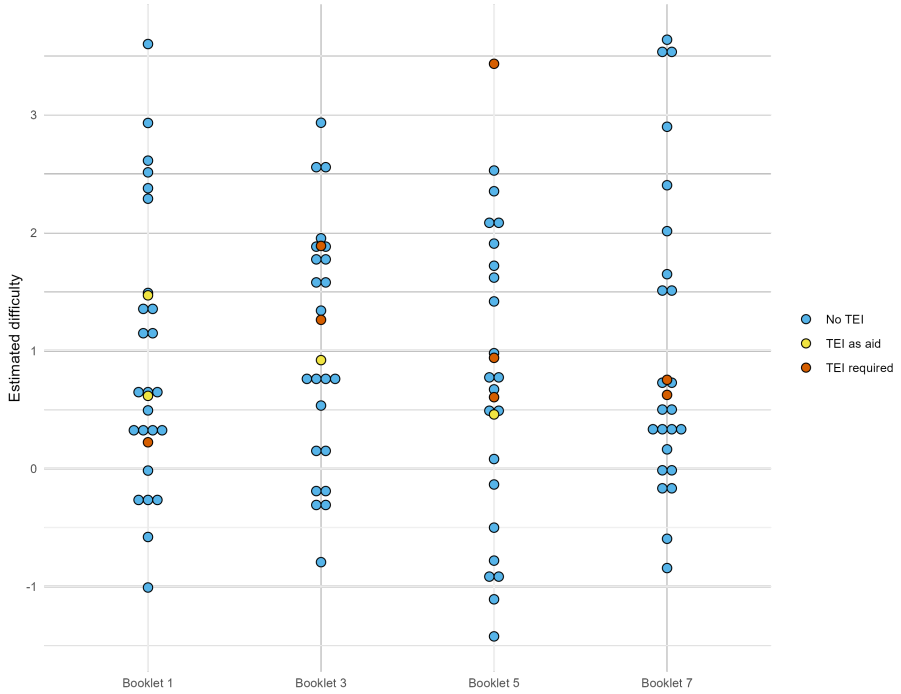


Figure 6.1: The estimated difficulty of TE items, as compared to the non-TE items.

Figure 6.2 shows the estimated difficulty of TE items is, as compared to the non-TE items for the PSI scale. Similar to Figure 6.1, the items for which technology was required are shown in orange, items for which technology could be used as aid are shown in in yellow. From Figure 6.2 it can be concluded that technology-enhanced PSI items are among the more easy items when compared to the other PSI items.

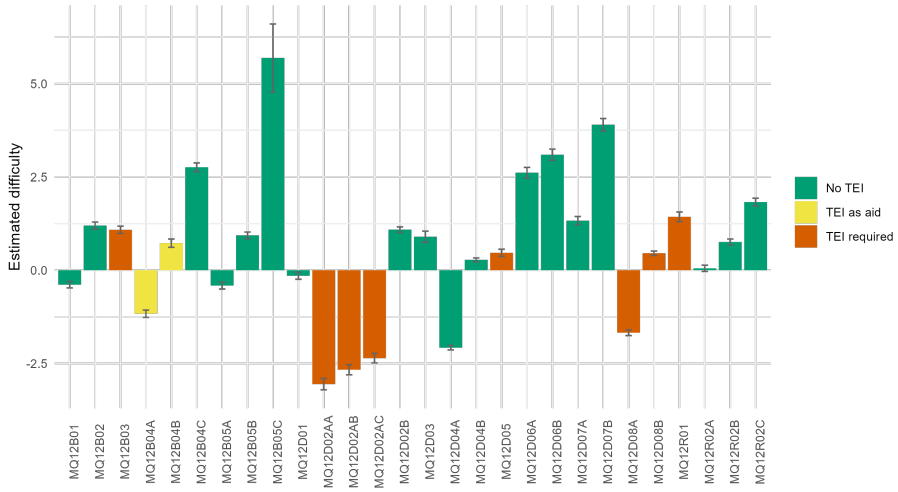


Figure 6.2: The estimated difficulty with their standard errors of TE items, compared to the non-TE items for PSI.

7. Study 6 – Process data

7.1 Aim Study 6

As discussed in the introduction section, the process data of technology-enhanced items in general, and the digitally-based interactive PSI items in particular potentially contain valuable information about inter-individual and intra-individual differences in response processes. In this study, we explored the potential of the process data of the eTIMSS and PSI scale to reveal differences in response strategies. Specifically, we conducted analyses of the on-screen time as a proxy of response time. First, using a DIF analysis, we compared the response times across technology-enhanced items and the more traditional items of the PSI scale and the eTIMSS scale. Next, we studied DIF in the time variables of the eTIMSS and PSI scales across countries. Consequently, we tested for systematic intra-individual differences in response processes on the PSI items, and we explored if response time can explain differences in item performance across countries. Finally, we conducted some univariate analyses on the use of the calculator for the water tank items.

7.2 Method Study 6

7.2.1 DIF of response times

We conducted a similar DIF analysis on the response times as discussed above for Study 5. That is, we fit a unidimensional linear factor model (i.e., similar to the model for the item scores above, but in this case for continuous variables; Molenaar et al., 2015), to the logarithmically transformed screen time variables of the PSI and eTIMSS scales, and tested for uniform DIF. We focused on the estimates of the technology-enhanced items and the non-enhanced items. In addition, we tested whether the time variables contain DIF when compared across countries. To be a meaningful source of information, the process data of the PSI and eTIMSS (i.e., time in this case), should capture the same individual differences across countries (i.e., not display a lot of DIF).

7.2.2 Intra-individual differences

To explore the value of the PSI process data as a source of intra-individual differences in response processes, we tested for local dependencies in the responses and the response time variables. That is, when controlling for individual differences in response accuracy and response speed, additional item-specific dependencies between response time and response accuracy indicate intra-individual differences in response processes (e.g., strategies, motivation, etc.; see Bolsinova et al., 2017). To this end, we applied the methodology by Ranger and Ortner (2012) which involves modeling the possible local dependencies using residual covariances. A significant residual covariance indicates that, after accounting for individual differences in speed and accuracy, there are item specific dependencies between speed and accuracy (e.g., for a given item, a slower response has a higher expected score).

7.2.3 Response processes to explaining item differences

To test if differences in item scores across countries can be explained in terms of differences in response processes, we applied the newly developed methodology by Molenaar and Feskens (2022, Manuscript in preparation). That is, we explored whether some of the DIF in item scores found in Study 5 across countries can be attributed to differences in response processes. Key idea of the method by Molenaar and Feskens is that item specific differences in difficulty across countries (partly) disappear once the process data is taken into account. This indicates that a difference in response process underlies the difference in item difficulty (e.g., in one country the respondents use a more efficient or less error-prone strategy to solve a given item). To this end, we focused on three comparisons: 1) England and USA; 2) Hong Kong and Malaysia; and 3) Sweden and Turkey. These countries were selected as they are assumed to be sufficiently similar to avoid uninterpretable results. In the method by Molenaar and Feskens (2022, Manuscript in preparation), it is tested if a full DIF effect can be (partly) explained by item specific intra-individual differences in the response times across countries. These DIF effects go beyond the overall differences across countries on accuracy (latent ability) and time (latent time).

7.3 Results Study 6

7.3.1 DIF of response time

To gain insight into whether the measurement of mathematics skill using technology-enhanced items is confounded by computer skills, technology-enhanced items were compared to non-TE items in terms of logarithm time and the number of countries in which a country is flagged as DIF with respect to the time spend on an item. Also, the pattern in the countries that showed DIF was investigated. The results are presented for eTIMSS items first, followed by the PSI items.

Response time of TE eTIMSS items

Figure 7.1 depicts the response times of TE items, in relation to non-technology-enhanced items in Booklet 1, 3, 5 and 7. From Figure 7.1 it can be concluded that no systematic differences regarding the response times of TE items are found.

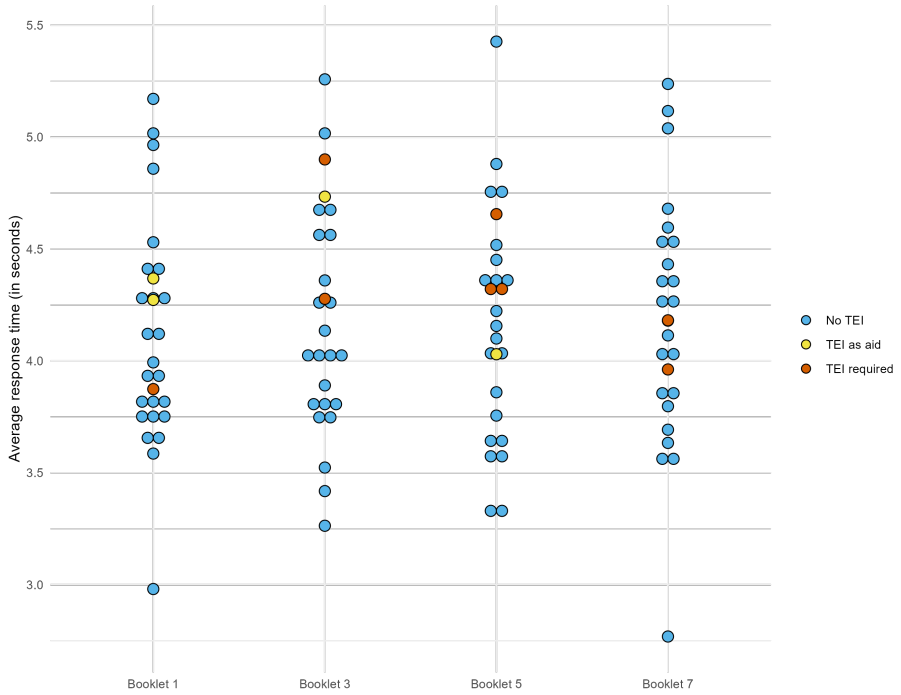


Figure 7.1: The estimated response times for TE items in eTIMSS booklets, compared to non-TE items.

Response time of TE PSI items

In Figure 7.2 the response times of the PSI screens are presented. The non-technology enhanced screens are presented in green, screens that included technology-enhanced items in orange and the technology as aid items in yellow. In cases where multiple items were presented on a screen, the response time on the screen was divided by the number of items, to present the 'item' time. Figure 7.2 illustrates that two technology-enhanced items required the longest time to response. This is MQ12B03, in which students had to draw the contour of three building elements. In MQ12R1, students were required to deduce the robots' rule by filling out numbers. These two items required students to perform multiple interactive acts, which could be an explanation for the longer response times. The other technology-enhanced items are comparable in terms of response times to the non-technology-enhanced items.

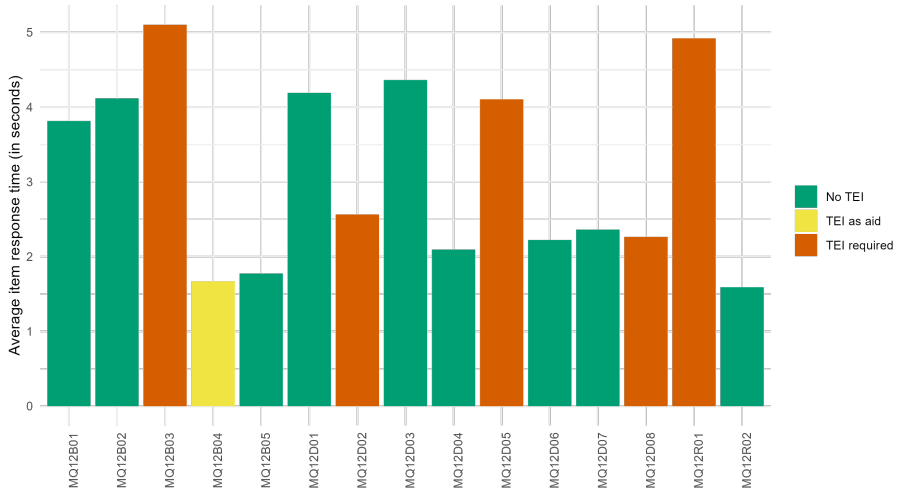


Figure 7.2: The estimated response times for TE items in PSI items, compared to non TE items.

DIF with respect to time across country

eTIMSS

Table 7.1 provides an overview of the number of items that show either a significantly longer response time (positive DIF) or significantly slower response time (negative DIF) on regular eTIMSS and technology-enhanced items (TEI). Note that if an item shows positive DIF in a given country, this means that the respondents in this country on average spent proportionally more time on this specific item as compared to other countries. See our discussion in the methods section for details.

Table 7.1: Overview of the number of items that show either a significantly longer response time (positive DIF) or significantly slower response time (negative DIF) on regular eTIMSS and technology-enhanced items (TEI).

	Regular items	**Required TEI**	**TEI as aid**
N-items	94	6	4
N-items with DIF	92	5	4
Mean number of countries positive DIF	3.0 (0-13)	2.3 (0-5)	3.5 (1-6)
Mean number of countries negative DIF	2.7 (0-11)	1.5 (0-3)	1.0 (0-2)

Table 7.1 shows that the range of the number of countries showing positive and negative DIF is within the range of the regular eTIMSS items.

PSI

Similar to Table 7.1, a comparison between non-technology enhanced PSI items and technology-enhanced PSI items is made with respect to items with DIF in Table 7.2.

Table 7.2: Overview of the number of items that show either a significantly longer response time (positive DIF) or significantly slower response time (negative DIF) on regular PSI and technology-enhanced (TE) screens.

	Regular PSI screens	**Required TE screens**	**TEI as aid screen**
N-items	9	5	1
N-items with DIF	9	5	1
Mean number of countries positive DIF	2.8 (0-7)	1.4 (0-3)	1
Mean number of countries negative DIF	2.7 (0-5)	3.2 (1-6)	2

Table 7.2 shows that on average, fewer countries have a significantly longer response time for technology-enhanced screens compared to regular PSI screens. Also more countries have significantly slower response times on technology-enhanced screens compared to regular PSI items. Nevertheless, the ranges are similar to each other.

7.3.2 Intra-individual differences

To start with, the correlation between individual differences in response time and response accuracy are equal to 0.487 (sd: 0.127), indicating that, in general, a higher score is associated with a slow response. Figure 7.3 depicts boxplots of the residual covariances for each item of the PSI scale across countries. These covariances indicate specific items for which respondents show intra-individual differences in response time and accuracy (i.e., effects that go beyond the individual differences above). These intra-individual differences indicate a switch in response strategy or response process. As can be seen, most residual covariances are positive, which indicates that respondents who did well on this item significantly slowed down as compared to their overall speed. Most notably, items 5 (MQ12B05), 14 (MQ12D07) and 15 (MQ12D08) show such an effect with a relatively large positive covariance. For item 5 (MQ12B05) and item 14 (MQ12D07) this may be due to these items being the most difficult items of the PSI scale (see Study 5). Respondents who did well on this item noticed the difficulty level and spent significantly more time on this item than expected on the basis of their overall speed. For item 15 (MQ12D08) the residual covariance may be due to this item being somewhat different in setup as compared to the other PSI items, thus requiring relatively more time for a correct answer. Important to note is that high residual correlations are generally highly variable across countries, indicating the heterogeneity of the speed-accuracy relation across countries.

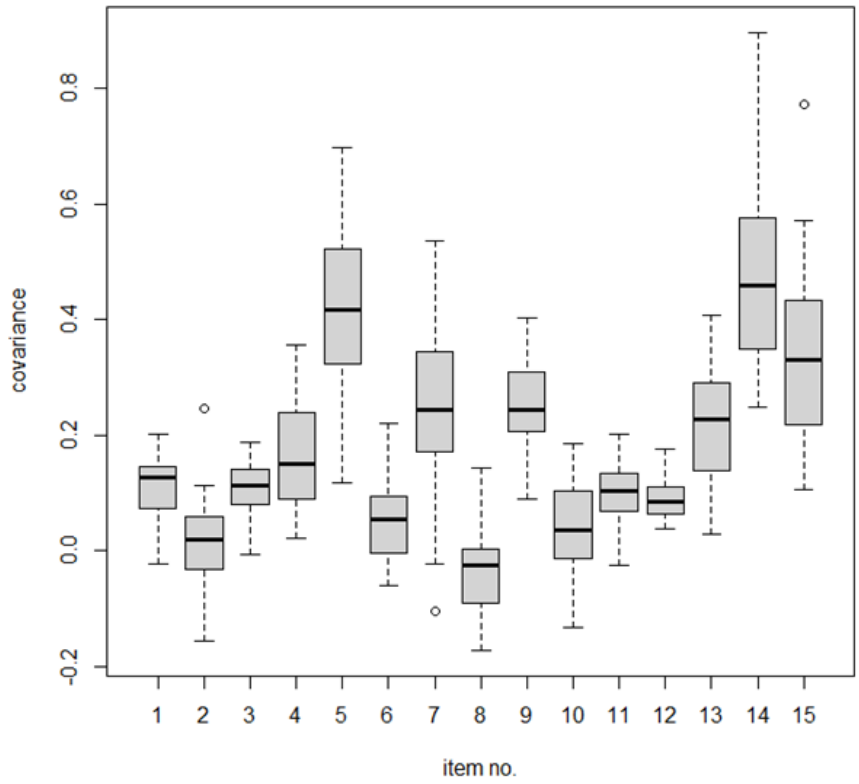


Figure 7.3: Boxplots of the residual covariances for each item of the PSI scale across countries. Item numbering is the same as in the table above.

7.3.3 Response processes to explaining item differences

As explained above, our goal was to test whether a full DIF effect can be (partly) explained by item specific intra-individual differences in the response times across countries, since these DIF effects go beyond the overall differences across countries (see Table 7.3 for the main effects of accuracy and time). As can be seen, England and USA do not differ significantly on overall accuracy, while USA on average takes more time. Hong Kong on average scores higher on overall accuracy as compared to Malaysia, while Malaysia on average takes more time. Finally, Sweden on average scores higher on overall accuracy as compared to Turkey, while Turkey on average takes more time.

Table 7.3: Regression of the overall accuracy (latent ability variable) and the overall time (latent time variable) on the country indicator variable for the different comparisons.

Comparison	**Overall accuracy**		**Overall time**	
	Estimate	*SE*	*Estimate*	*SE*
England – USA	-0.018	0.058	0.232*	0.059
Hong Kong – Malaysia	-0.915*	0.064	1.098*	0.063
Sweden – Turkey	-0.451*	0.067	0.477 *	0.065

If we look beyond these main effects, there are item specific differences in accuracy (which indicates DIF of the responses as studied in Study 5) and time (as indicated by the DIF on the time variables as studied in the present study above). Using the approach discussed above, we investigated if these item specific differences in time can account for the item specific differences in accuracy (DIF). To this end, the full DIF effect was decomposed into an unexplained effect and an effect that can be accounted for by differences in response times. See Table 7.4 for the results.

Table 7.4: Decomposition of the Full DIF effect into an unexplained part, and a part that can be explained by differences in response times.

Item	Full DIF effect		Unexplained		Explained by time		Time effect	
	Estimate	*SE*	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
				England - USA				
MQ12B05 (Water tank)	0.333*	0.047	0.136	0.054	0.197*	0.045	0.327*	0.074
MQ12D01 (Foot length)	-0.520*	0.056	-0.185	0.074	-0.335*	0.053	0.677*	0.112
MQ12D05 (Stride length)	-0.423*	0.065	-0.492*	0.073	0.069	0.048	0.145	0.102
				Hong Kong - Malaysia				
MQ12B01 (Building size)	0.693*	0.067	-0.071	0.073	0.764*	0.052	2.764*	0.493
MQ12B03 (Constructing walls)	0.357*	0.065	-0.096	0.078	0.453*	0.052	1.594*	0.421
MQ12D02 (Foot Length)	0.255*	0.050	0.233*	0.063	0.021	0.048	0.046	0.104
MQ12D08 (Run race)	0.250*	0.060	0.457*	0.074	-0.207*	0.052	-0.421*	0.109
				Sweden - Turkey				
MQ12B03 (Constructing Walls)	-0.254*	0.077	-0.570*	0.093	0.315*	0.073	0.561*	0.142
MQ12R02 (Robots)	0.424*	0.063	0.172	0.084	0.253*	0.071	0.652*	0.205

The parameters are estimated for each country pair. However, as 'country' is being dummy coded into 0 and 1, the estimates reflect the difference in difficulty in the country coded as 1 relative to the difficulty in the country coded as 0. The country mentioned first is coded 0, see the main text.

For England (coded 0) and the USA (coded 1), three DIF items are identified on the basis of the results from Study 5 (see Appendix Table A.5). Note that in Study 5, all countries were being compared to an overall aggregate. Therefore, some items may demonstrate DIF in the overall analysis but not if specifically compared between two countries). The full DIF effect refers to the regression coefficient of the country variable in the regression of the item scores on the latent variable and the country variable. As USA was coded as 1, the full DIF effects indicate that for respondents from the USA item MQ12B05 was relatively easier than for respondents from England. For the other DIF items, the effect is reversed, with the items being more difficult for respondents in the USA as compared to respondents from England. From the 'time effects' column in the table, it can be concluded that for items MQ12B05 and MQ12D01, the USA responded significantly slower on these items compared to the overall response speed of England and the USA (i.e., the time effect is significant and positive for these items). For item MQ12D05, there are no item specific differences in time across the countries.

From the explained and unexplained effects, it can be seen that for item MQ12B05 and MQ12D01, all DIF can be explained using the response times. That is, the unexplained effect is insignificant for both items. For item MQ12B05, the estimate of the explained effect is positive, which indicates that the response time difference (as reported under the time effects column in the table) is associated with an increased accuracy on this item for the USA. Thus, the USA sample responded significantly slower to this item, with a significantly higher accuracy. For item MQ12D01, this effect is reversed. That is, due to the negative estimate of the explained effect, it can be concluded that the slower response time in the USA sample is associated with a smaller accuracy. Finally, for item MQ12D05, the DIF cannot be explained by differences in response times.

For the Hong Kong (coded as 0) – Malaysia (coded as 1) analysis, we see that all four items considered were easier for the Malaysia sample as compared to the Hong Kong sample. For item MQ12D08, the Malaysia sample responded faster as compared to the Hong Kong sample, for the other items, Malaysia was slower as compared to Hong Kong (see the time effects column in the table). For items MQ12B01 and MQ12B03, the full DIF effect can significantly be explained in terms of item specific differences on the response times, with the unexplained effect being insignificant. For both items, the explained effect is positive, indicating that the Malaysia sample responded significantly slower with a significantly higher accuracy. For item MQ12D08, the response times significantly account for some part of the effect, but a significant unexplained effect remains. As a result, the effect is difficult to interpret. For item MQ12D02, the DIF cannot be explained in terms of differences in response times.

Finally, for the Sweden (coded as 0) and Turkey (coded as 1) comparison, we see that three items are considered. The difficulty of the first item, MQ12B03, is larger in the Turkey sample but smaller in the other two items. In addition, for items MQ12B03 and MQ12R02, Turkey responded slower as compared to Sweden as compared to the overall difference in response speed. For item MQ12D06, there is no item specific difference in time.

With respect to the DIF, for item MQ12B03, the DIF effect can be partly explained by the response times but a significant unexplained effect remains, which makes the effect difficult to interpret. For item MQ12R02, all DIF is explained by the response times, indicating that the slower response speed in the Turkey sample is associated with larger accuracy on this item. For the third item considered, MQ12D06, the DIF cannot be explained in terms of differences in response times.

If we compare the results in Table 7.4 across the different country comparisons, we see that results are heterogeneous across countries, even if a comparison involves the same item. That is, the DIF on item MQ12B03 can be fully explained by the response times in the Hong Kong – Malaysia comparison, but it contains a substantial unexplained part in the Sweden – Turkey comparison. In addition, for some items, taking more time increases the accuracy (and explains positive DIF) while for other items taking more time decreases the accuracy (and explains negative DIF), even in the same country comparison.

Calculator use

To explore if the use of the calculator by the respondents indicates a difference in solution strategy, we compared the mean accuracy and mean response time of the water tank items (MQ12B05A and MQ12B05B) across subjects who did and did not use the calculator. See Figure 7.7 for the results. In Figure 7.4 the mean item score on item MQ12B05A for respondents who did or did not use the calculator across countries are displayed. In Figure 7.5 the mean item score on item MQ12B05B for respondents who did or did not use the calculator across countries can be found. Finally, the response time across countries for respondents who used the calculator on both items, only on one of the items, or on none of the items are displayed in Figure 7.6. Countries are ordered according to the mean scores of the respondents who did not use a calculator on item MQ12B05A (i.e., the black line in Figure 7.4).

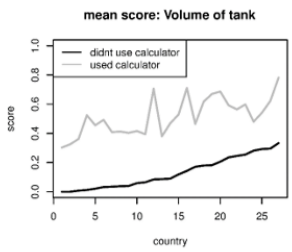


Figure 7.4: Mean item score on item MQ12B05A

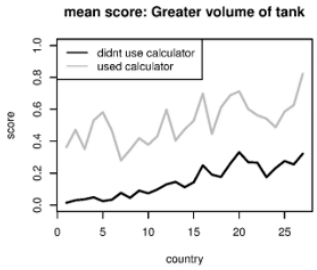


Figure 7.5: Mean item score on item MQ12B05B

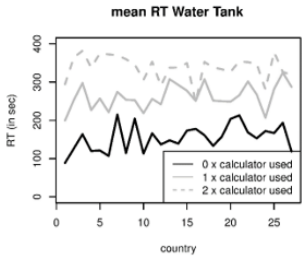


Figure 7.6: Mean response time of the water tank items

Figure 7.7: Mean item score and response time water tank item.

As can be seen, respondents who used the calculator (on one or both of the items) took more time but had a substantial higher mean score on the items as compared to respondents who did not use the calculator. These differences are stable across countries. The respondents who did not use the calculator took 153.85 seconds averaged over countries (SD: 34.37), which indicates that most of these respondents did put serious effort in the item, but either failed to see the possibility of the calculator, or did not think a calculator was necessary. In Figure 7.8, the proportion of respondents who used the calculator is plotted against the mean score on the two items across countries. The figure depicts the proportion of respondents who used the calculator in a given country (x-axis) and the mean score in that country (y-axis). As can be seen, there is a strong linear relation between these two variables (correlations are 0.763 and 0.764 for items MQ12B05A and MQ12B05B respectively), meaning that countries in which more respondents used the calculator generally scored higher on these items. Therefore, it may be that countries differ in how common it is to use a calculator during arithmetic questions. This should be taken into account when interpreting the differences across countries on questions that enable the use of the calculator.

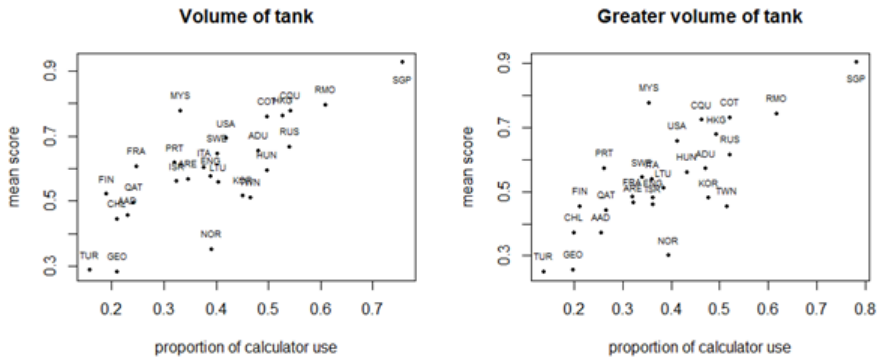


Figure 7.8: Scatterplot of calculator use and mean item score for the different countries.

8. Study 7 - Exploring response behavior

8.1 Introduction

One of the primary objectives of TIMSS is to compare educational outcomes among all participating countries. To ensure that valid statements can be made about the average proficiency of a group, it is crucial that the test is comparable across all countries in both assessments. When the conditions hold such that scores are comparable across groups, this is called *measurement invariance* (Millsap, 2011). If a test is not measurement invariant, it means that at least one item functions differently across groups, which is referred to as *differential item functioning* (DIF). Apart from score comparability, equal functioning across demographic variables is also important to ensure fairness of testing. DIF may occur due to differences in language, culture and education across countries. Traditionally, DIF has been considered a nuisance factor and, therefore, as a threat to validity. However, recent studies propose a different perspective, suggesting that DIF may actually provide valuable information about the differences between countries, highlighting the weaknesses and strengths of their educational systems (Cuellar et al., 2021; Verhelst, 2012). Within this study we follow this line of reasoning in the evaluation of the PSI and eTIMSS assessment.

Additionally, we employ cluster techniques to uncover valuable information from the process data obtained in the digital assessment, following a similar approach as proposed by Salles and colleagues (Salles et al., 2020).

In both approaches, we will follow a similar procedure. First, we will analyze the PSI and eTIMSS assessments separately, only including those countries in the eTIMSS analysis that have also participated in the PSI assessment. These results aim to extract meaningful didactic information from the data. Second, we will compare the outcomes of the PSI and eTIMSS assessments, which can contribute to building the validity argument for both assessments.

8.2 Methods

8.2.1 An alternative view on DIF

Rather than considering differential item functioning (DIF) as an individual item property, Bechger and Maris (2015) argue that it might be more sensible to define DIF in terms of pairs of items. Unlike traditional procedures, DIF is defined in terms of the relative difficulties of pairs of items (which are identified from the observations) and not in terms of the difficulties of individual items. This procedure starts with a separate calibration of the data within each group. Building upon Bechger and Maris's study, Cuellar et al. (2021) have extended the procedure to accommodate multiple groups as well. In our analysis, we will adhere to the guidelines outlined in their study.

In this study, the item parameters are estimated separately for each country. Once the item parameters have been estimated, the next step is to summarize the results. In their work, Cuellar et al. have explored different multivariate analysis techniques with an emphasis on visualization to evaluate DIF. The key concept is that all relevant information can be captured in the (group-specific) distance matrix between item difficulties[1]. The analysis of DIF can consequently be seen as a form of residual analysis after the dominant component has been removed. Singular value decomposition (SVD) (Eckart & Young, 1936) can be used to obtain the residual structure.

In the analysis, the first step is to identify poorly functioning items within each participating country. This is achieved by computing the p-value (or mean item score) for each item in each country. It was found that item M72017 showed no score variation in Georgia within the eTIMSS assessment, leading to its exclusion from further analysis. Similarly, since there is only one item (MQ12D01) measuring the content domain of numbers in the PSI assessment, it has also been removed from further analysis.

The subsequent step involves computing a Rasch model within each country separately. To conduct these analyses, we utilized the R package *dexter* (Maris et al., 2022). Next, the item content domains have been matched to the estimated item parameters.

We used SDV to estimate various dimensions in the data and removed the first component as we are interested in DIF, not ability. After obtaining the data matrix, we visualized the results using a heatmap. Heatmaps offer an ordered representation of the data matrix using a color scale. In this study, we generated the heatmaps using the R package *pheatmap* (Kolde, 2019).

It is also possible to use various options in the functions that produce the heatmap. One can focus on evaluating the structure of the countries on the one hand or alternatively focus on evaluating the structure of the items. The former utilizes Euclidean distances and complete linkage (Everitt et al., 2011), whereas the latter employs average linkage (Hastie et al., 2009), see also Cuellar et al., 2021.

[1] https://dexter-psychometrics.github.io/

8.2.2 Clustering process data

Digital assessment in general and the use of technology-enhanced items in particular provide the opportunity to capture additional log or process data. Process data is increasingly deliberately captured by design - rather than considered a mere by-product - in order to obtain a better understanding of test-taker performance and engagement (Maddox, 2023). The field of educational data mining (EDM) is one of the methodologies that can be applied to evaluate process data (Salles et al., 2020). Romero and Ventura define EDM as using computational approaches to analyze educational data in order to address educational inquiries (Romero & Ventura, 2010).

Among the various techniques available, unsupervised learning techniques are particularly useful for identifying meaningful patterns in process data. In line with the approach taken by Salles et al., 2020 we have applied Density Based Clustering of Applications with Noise (*dbscan*) (Hahsler et al., 2019) to cluster the countries within the PSI and eTIMSS assessments. In this analysis, we have utilized item scores and response times collected from both assessments. Although we have explored the inclusion of other process data, it did not yield meaningful outcomes.

8.3 Results

In Paragraph 8.3.1, we will present the findings of the measurement invariance analysis. Following that, in Paragraph 8.3.2, we will discuss the results obtained from the cluster analysis utilizing process data.

8.3.1 Differential item functioning

PSI

For the PSI assessment, Figures 8.1 and 8.2 display the cluster country and cluster item structure respectively.

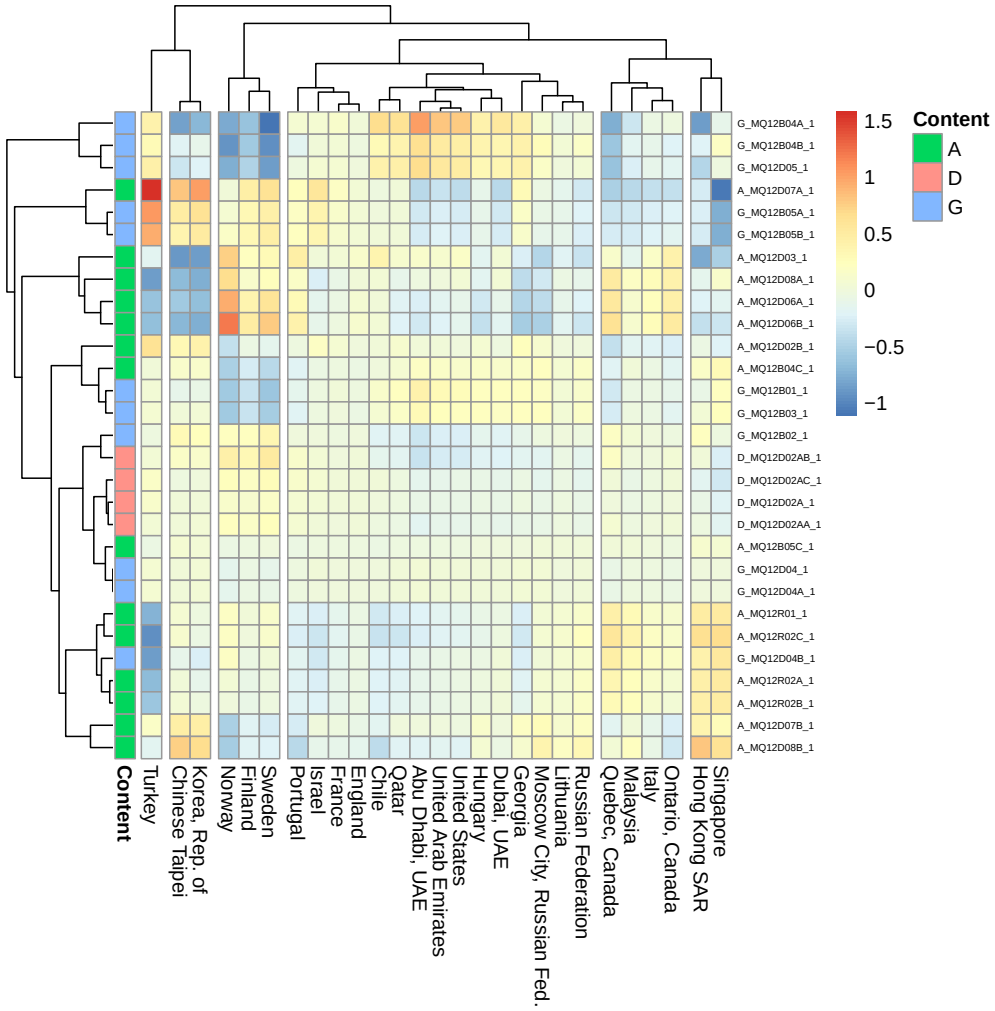


Figure 8.1: Country structure PSI assessment.

Figure 8.1 presents the separately estimated item difficulty parameters for each country. These values have been used to cluster the countries based on their similarities. Six distinct clusters have been identified. The content domain of the items is indicated by the bar on the left side of the figure.

Within each cluster, subtle relative differences in item difficulty can be observed. Items that are colored blue indicate better performance by countries on those specific items. For instance, Korea and Chinese Taipei, belonging to the second cluster, have demonstrated relatively strong performance on the first three items, all measuring Geometry. This pattern is even more pronounced for the three Scandinavian countries (Norway, Finland, and Sweden) clustered together, as indicated by the dark blue color. Turkey, on the other hand, exhibits a somewhat distinctive response pattern and is identified as a separate cluster. It shows relatively strong performance in the content domain of Algebra. Hong Kong and Singapore form a distinct cluster and are not clustered with other East Asian entities such as Chinese Taipei and Hong Kong. Lastly, a larger cluster is observed, characterized by a relatively weaker performance in Geometry and a somewhat stronger performance in Data and probability.

Figure 8.2 depicts the same results as Figure 8.1, which showcase the item difficulty

parameters. However, in Figure 8.2, the cluster analysis arranges these values in a different manner. This alternative arrangement offers the potential for greater insight into the item structure in the PSI assessment.

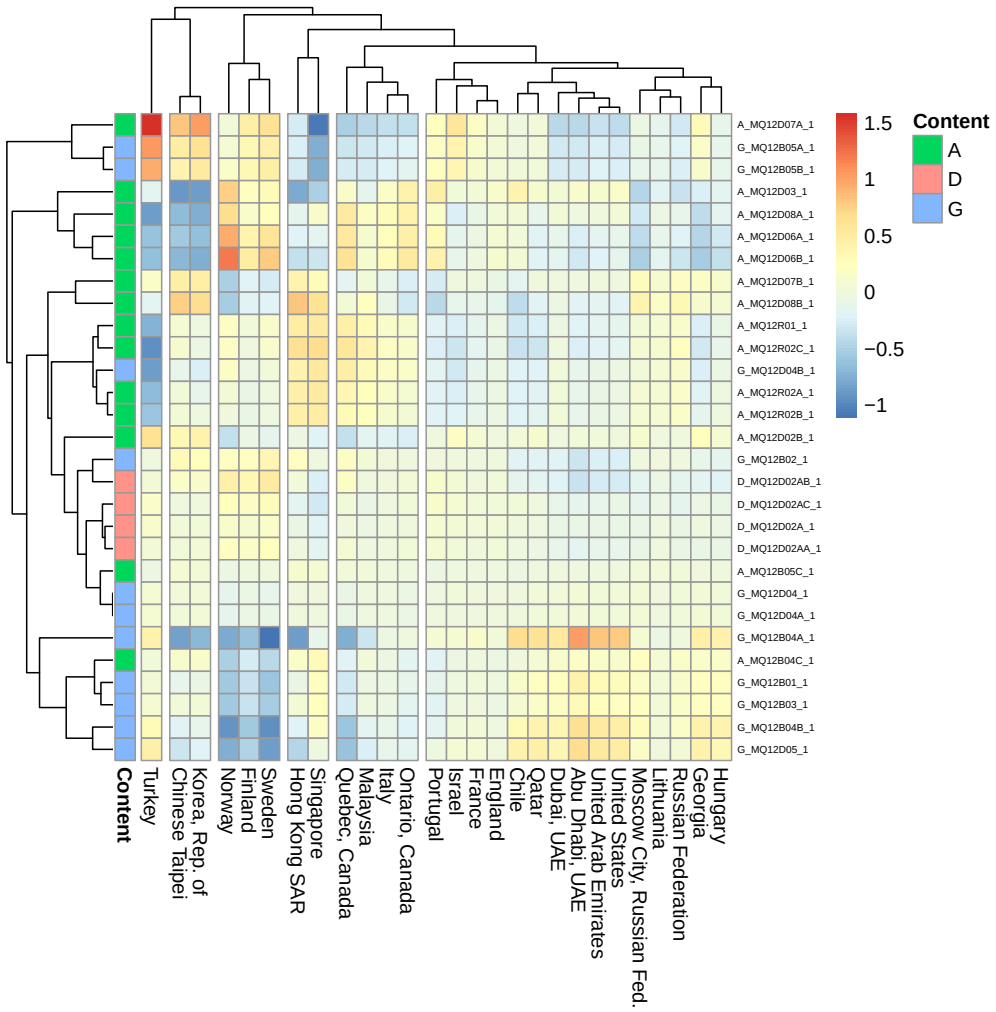


Figure 8.2: Item structure PSI assessment.

The first thing that is noticeable from Figure 8.2 is that the content domains are quite well identified as distinct clusters. Moreover, it becomes even more evident that Scandinavian countries exhibit relatively strong performance in the Geometry content domain while displaying comparatively weaker performance in Algebra. Conversely, Turkey demonstrates the opposite pattern, with relatively strong performance in algebra and comparatively weaker performance in geometry.

eTIMSS

To assess the comparability between the eTIMSS and PSI assessments, we have conducted a parallel analysis for the eTIMSS data. For this analysis, we have included only those countries that have participated in both the eTIMSS and PSI assessments. By comparing the relative performances of countries within both assessments, we aim to validate the extent to which country performances differ between eTIMSS and PSI. This comparison

allows us to gauge the extent to which PSI and eTIMSS measure different competences and consequently to what extent PSI and eTIMSS measure different competences.

In Figures 8.3 and 8.4 the cluster analysis results for the eTIMSS assessment can be found.

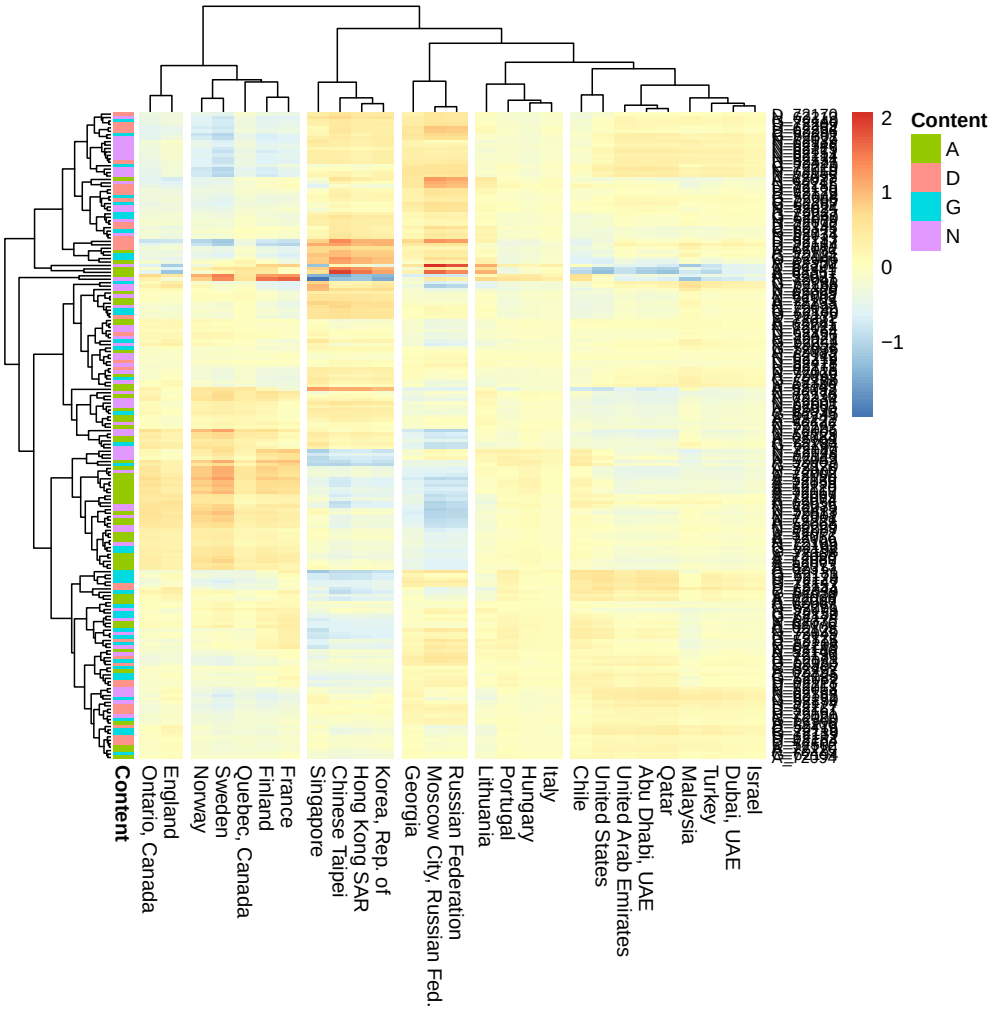


Figure 8.3: Country structure eTIMSS assessment.

Figure 8.3 presents the results of the differential item functioning (DIF) analysis conducted for the eTIMSS assessment. Interpreting the plot is more challenging due to the larger number of items included compared to the analysis based on the PSI assessment. However, several interesting patterns can still be observed.

Once again, the Scandinavian countries are clustered together, but this time, Quebec and France are also included in this cluster. Countries within this cluster demonstrate a relatively strong performance in the Numbers domain, as indicated by the light blue color in the upper part of the graph. As aforementioned, the Numbers domain was not included in the PSI analysis, so it is not surprising that the allocation of countries to clusters is somewhat different in analysis of the eTIMSS assessment.

Another interesting finding is that Singapore, Chinese Taipei, Hong Kong, and Korea are clustered together in contrast to the PSI assessment. This suggests that the competences

required by the PSI assessment differ somewhat from those measured in the eTIMSS assessment. Determining whether this discrepancy is due to measurement noise or represents a meaningful signal is to decide by substantive experts.

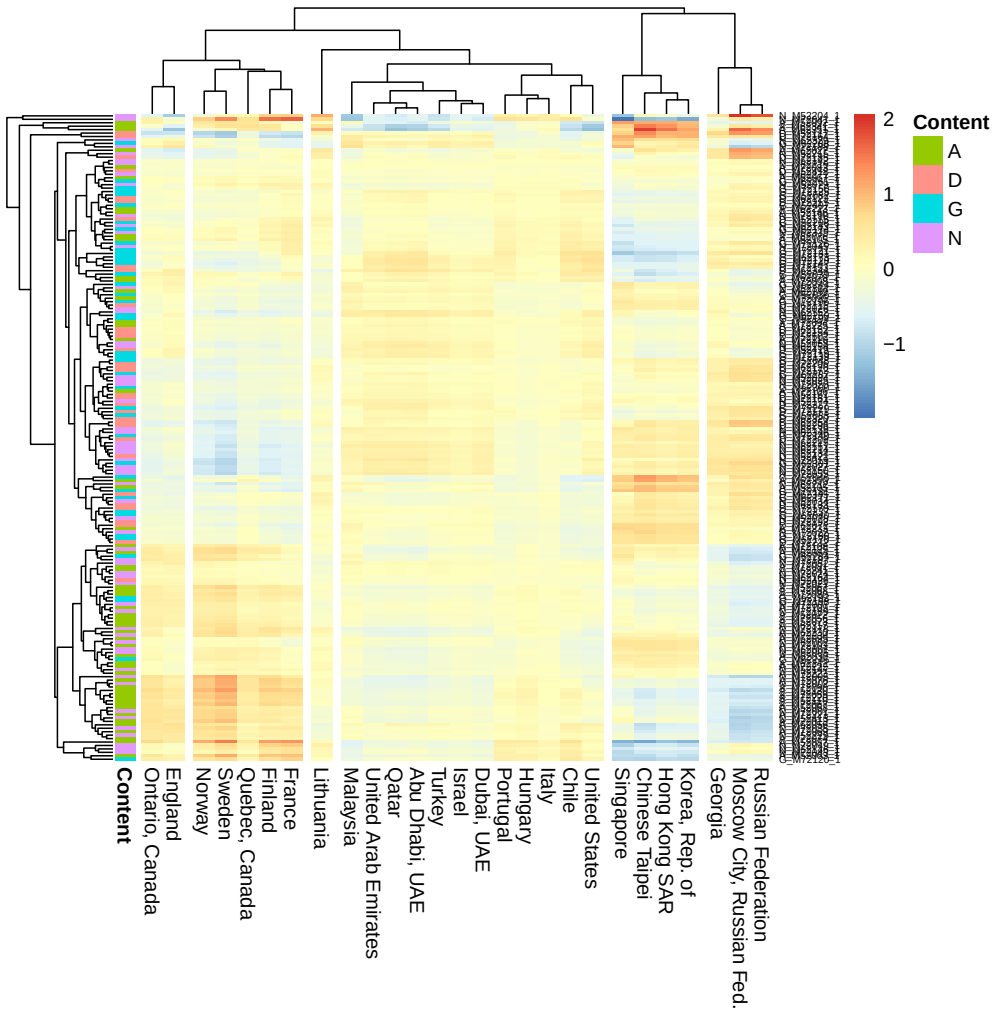


Figure 8.4: Item structure eTIMSS assessment.

Although somewhat less clear, it can be seen in Figure 8.4 that in the eTIMSS assessment as well the items are reasonably well clustered based on their underlying content domains. The first two clusters (Ontario and England; Norway, Sweden, Quebec, Finland and France) exhibit relatively weaker performance on items measuring algebra. In contrast, the countries in clusters five and six (Singapore, Chinese Taipei, Hong Kong, Korea and Georgia, Moscow and the Russian Federation respectively) demonstrate relatively stronger performance on these algebra items.

8.3.2 Process data

PSI

Within Figure 8.5 the cluster analysis results for the PSI assessment are displayed.

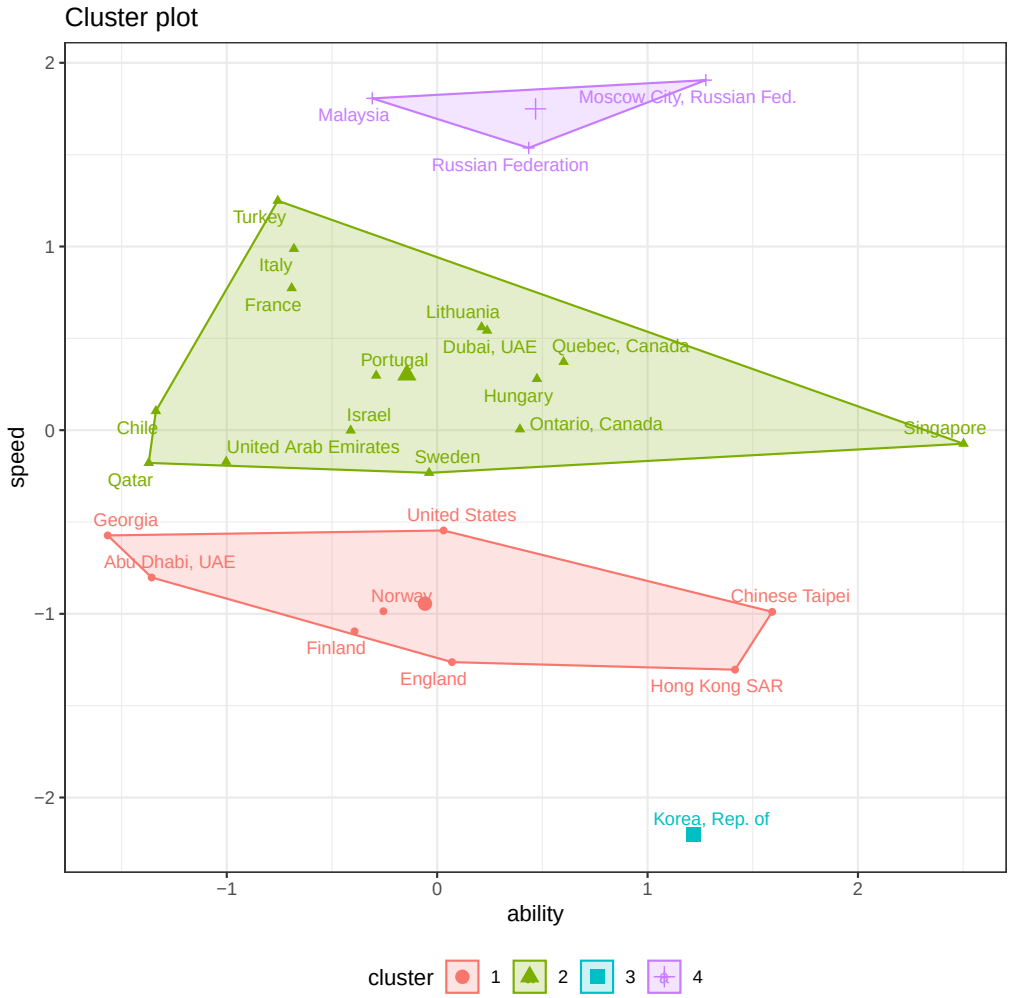


Figure 8.5: Cluster analysis PSI.

Figure 8.5 reveals the identification of four distinct clusters based on the response patterns in the PSI assessment, utilizing the dbscan algorithm (Hahsler et al., 2019). Notably, Korea forms a separate cluster, characterized by its unique response pattern. Korean students exhibit exceptional speed in responding and demonstrate above-average proficiency. On the other hand, cluster four comprises students who require more time to complete the PSI items. Within this cluster, we find Malaysia, Moscow, and the Russian Federation.

eTIMSS

Figure 8.6 displays the cluster outcomes for eTIMSS. Once again Korea forms a separate cluster. Here, Malaysia is also a cluster on its own. Overall, however, the cluster results based on the response behavior - both item scores and item response times - are very similar for the PSI and eTIMSS assessment.

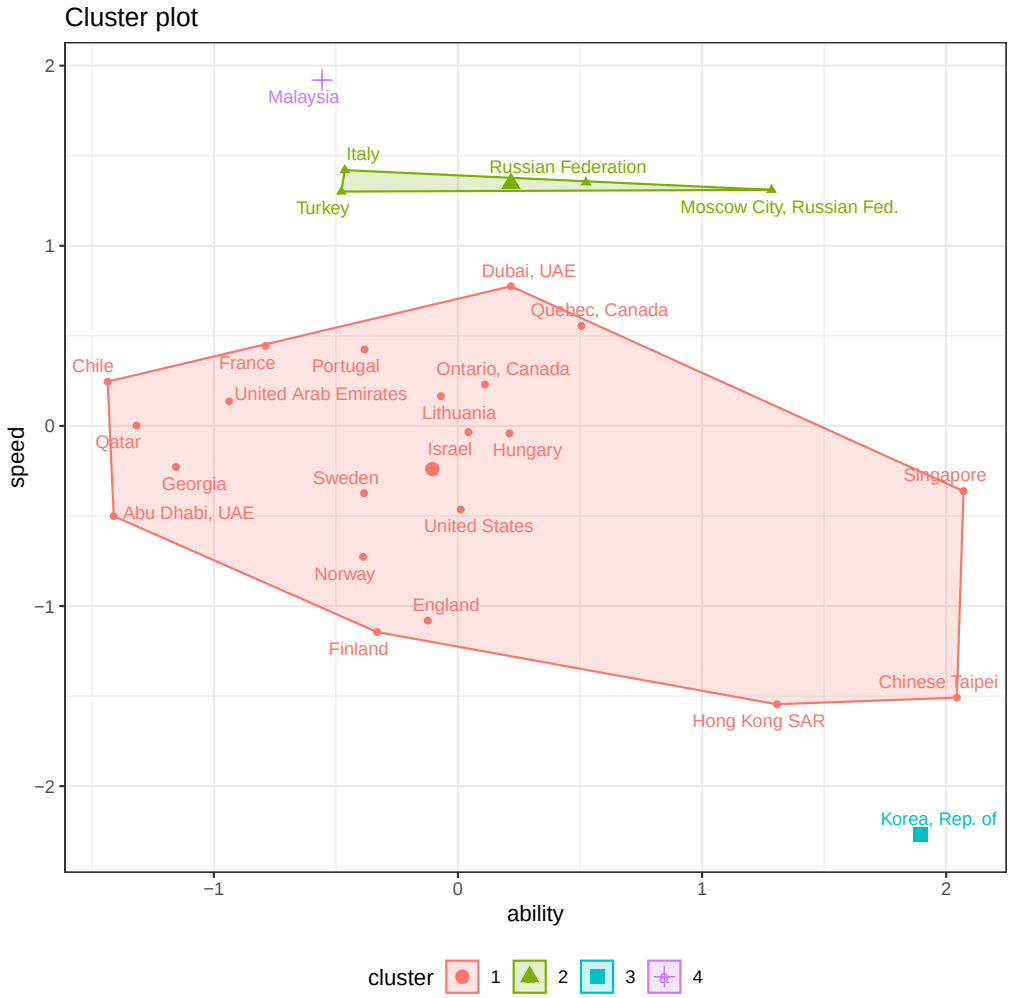


Figure 8.6: Cluster analysis eTIMSS.

8.4 Conclusion

In this supplement study, we have conducted an evaluation and comparison of the item and country structures within the PSI and eTIMSS assessments. To accomplish this, we utilized measurement invariance techniques and cluster analysis. Notably, the measurement invariance techniques yielded intriguing findings.

Differences found in, for example, the clustering of Asian countries could be used as a starting hypothesis for further research by substantive experts. Overall, the country and item structure found in the PSI and eTIMSS assessment are - although not entirely - comparable, indicating that both assessments call upon the same competences. These subtle differences found between both assessments can open avenues for in-depth research.

A. Appendix

A.1 Tables

Table A.2: eTIMSS Booklet 3: Results from the classical DIF study 5 on the responses of the eTIMSS scale items.

Item ID	Count	Countries
M62005	5	FIN, FRA, MYS, PRT, ENG
M62027	11	CHL, FRA, GEO, HKG, HUN, ITA, LTU, PRT, RMO, RUS, TUR
M62084	2	GEO, RMO
M62132	1	MYS
M62132	1	HUN
M62139	4	FIN, NOR, SWE, COT
M62142	1	FRA
M62164	1	TUR
M62174	3	HKG, KOR, MYS
M62223	3	HUN, SWE, USA
M62244	5	ISR, PRT, RMO, RUS, TUR
M62254	8	FRA, HUN, KOR, RMO, SGP, SWE, TUR, COT
M62261	2	KOR, AAD
M62300	2	ITA, SWE
M62300	5	TWN, HUN, ISR, ITA, SWE
M62351	16	CHL, HUN, ISR, MYS, NOR, PRT, QAT, SGP, SWE, ADU, AAD, ARE, TUR, USA, COT, ENG
M72020	2	FIN, TUR
M72020	3	FIN, RUS, TUR
M72027	10	CHL, FIN, FRA, HUN, NOR, PRT, SWE, TUR, COT, ENG
M72052	8	TWN, FRA, KOR, PRT, RUS, SGP, ADU, ARE
M72067	10	FIN, FRA, ITA, MYS, NOR, PRT, SWE, TUR, COT, CQU
M72083	0	
M72083	0	
M72108	8	GEO, ITA, LTU, MYS, NOR, QAT, RUS, SWE
M72108	5	FIN, MYS, NOR, SWE, ADU
M72126	3	TWN, HKG, SGP
M72126	4	TWN, HKG, KOR, SGP
M72164	9	FRA, HUN, ITA, KOR, RMO, SWE, USA, CQU, ENG
M72178	13	CHL, ITA, MYS, PRT, QAT, RUS, SGP, ADU, AAD, ARE, TUR, USA, COT
M72181	9	CHL, FIN, HUN, NOR, SGP, ARE, USA, COT, ENG
M72185	7	KOR, NOR, SGP, TUR, USA, COT, ENG
M72185	11	KOR, NOR, SGP, ADU, AAD, ARE, TUR, USA, COT, CQU, ENG
M72234	3	ADU, AAD, ARE

Note: 'Count' indicates the number of countries in which the item is flagged as DIF item.

Table A.3: eTIMSS Booklet 5: Results from the classical DIF study 5 on the responses of the eTIMSS scale items.

Item ID	Count	Countries
M52034	3	TWN, HUN, RUS
M52036	2	CHL, PRT
M52073	9	GEO, HKG, KOR, QAT, RMO, RUS, AAD, ARE, USA
M52078	3	TWN, HUN, SGP
M52105	12	FIN, FRA, HUN, NOR, PRT, SWE, ADU, AAD, ARE, TUR, USA, ENG
M52110	9	FIN, HKG, QAT, RMO, RUS, ADU, AAD, ARE, USA
M52117	8	TWN, SWE, ADU, AAD, ARE, COT, CQU, ENG,
M52130	10	CHL, FIN, FRA, ISR, NOR, RMO, RUS, SGP, COT, ENG
M52134	8	TWN, FIN, HUN, MYS, NOR, SGP, SWE, ENG
M52174	4	CHL, FIN, NOR, SWE
M52174	4	CHL, KOR, NOR, SWE
M52407	2	NOR, ENG
M52413	7	GEO, MYS, NOR, RMO, RUS, SGP, SWE
M52426	2	MYS, RMO
M52502	1	MYS
M62002	2	ITA, MYS
M62040	1	TUR
M62105	2	AAD, ARE
M62105	1	ARE
M62123	1	NOR
M62123	0	
M62133	3	HKG, KOR, TUR
M62149	0	
M62150	8	TWN, FIN, HUN, NOR, RUS, SGP, SWE, ENG
M62173	3	HUN, KOR, SGP
M62219	2	CHL, MYS
M62241	5	GEO, HUN, ISR, MYS, RUS
M62288	2	HUN, RUS
M62288	2	HUN, RUS
M62335	2	FIN, SWE

Note: 'Count' indicates the number of countries in which the item is flagged as DIF item.

Table A.4: eTIMSS Booklet 7: Results from the classical DIF study 5 on the responses of the eTIMSS scale items.

Item ID	Count	Countries
M52039	3	ISR, KOR, ENG
M52048	8	TWN, FRA, HUN, ISR, LTU, PRT, COT, ENG
M52067	3	MYS, NOR, SWE
M52068	2	FRA, RMO
M52079	5	ISR, ITA, PRT, TUR, CQU
M52087	14	TWN, GEO, HKG, ISR, KOR, QAT, RMO, RUS, SGP, ADU, AAD, ARE, TUR, USA
M52115	3	FIN, FRA, RUS
M52147	1	MYS
M52204	14	FIN, FRA, GEO, HUN, ISR, LTU, MYS, NOR, RMO, RUS, TUR, USA, COT, ENG
M52208	5	MYS, QAT, SGP, AAD, ARE
M52215	9	TWN, FIN, GEO, HUN, LTU, PRT, RMO, RUS, SWE
M52364	8	TWN, FRA, HKG, ISR, MYS, SGP, ADU, ARE
M52419	7	GEO, ISR, ITA, ADU, AAD, ARE, TUR
M52419	11	CHL, FIN, FRA, HKG, HUN, LTU, PRT, RMO, SWE, USA, ENG
M52421	12	FRA, HKG, ISR, ITA, LTU, MYS, NOR, QAT, RUS, SWE, AAD, ARE
M72002	7	FIN, FRA, HUN, NOR, RUS, SWE, TUR
M72035	10	TWN, GEO, ITA, KOR, RMO, RUS, ADU, AAD, ARE, TUR
M72055	5	FIN, NOR, SWE, USA, CQU
M72090	10	GEO, ISR, MYS, QAT, RMO, RUS, ADU, AAD, ARE, TUR
M72106	2	CHL, GEO,
M72106	3	TWN, SWE, TUR
M72106	7	ISR, PRT, QAT, SWE, ARE, TUR, COT
M72119	4	FRA, HUN, KOR, MYS
M72128	6	FIN, HKG, ITA, KOR, COT, CQU
M72128	3	FIN, ITA, USA
M72128	3	HKG, ITA, RUS
M72153	3	RUS, AAD, ARE
M72153	6	HKG, HUN, QAT, SGP, ADU, CQU
M72172	8	CHL, FIN, FRA, KOR, LTU, SWE, TUR, USA
M72188	10	CHL, FRA, GEO, HUN, ITA, MYS, RMO, RUS, USA, CQU
M72222	4	HUN, ITA, KOR, USA
M72233	3	CHL, FIN, NOR

Note: 'Count' indicates the number of countries in which the item is flagged as DIF item.

Table A.1: eTIMSS Booklet 1: Results from the classical DIF study 5 on the responses of the eTIMSS scale items.

Item ID	Count	Countries
M52024	3	FIN, ISR, NOR
M52046	16	CHL, FIN, FRA, HUN, MYS, NOR, PRT, QAT, ADU, AAD, ARE, TUR, USA, COT, CQU, ENG
M52058	3	FIN, AAD, ARE
M52058	6	CHL, NOR, SWE, ADU, USA, CQU
M52063	14	CHL, FIN, FRA, HUN, ITA, LTU, NOR, PRT, SWE, ARE, TUR, COT, CQU, ENG
M52072	4	ITA, RMO, SWE, COT
M52082	5	TWN, HUN, KOR, MYS, USA
M52083	4	GEO, HKG, SWE, TUR
M52092	11	CHL, GEO, ISR, QAT, RMO, RUS, ADU, AAD, ARE, TUR, USA
M52125	9	CHL, ITA, LTU, NOR, PRT, RMO, RUS, SWE, TUR
M52146	5	CHL, TWN, MYS, PRT, USA
M52146	7	FIN, HUN, LTU, PRT, SWE, ADU, COT
M52161	11	GEO, HKG, MYS, NOR, PRT, QAT, RUS, SGP, ADU, AAD, ARE
M52229	14	TWN, FIN, GEO, ITA, KOR, LTU, QAT, RMO, RUS, SGP, ADU, AAD, ARE, TUR
M52418	1	HUN
M52418	1	MYS
M72007	4	SWE, AAD, TUR, ENG
M72007	5	GEO, ITA, ADU, AAD, ARE
M72017	3	TWN, GEO, KOR
M72025	7	FIN, ISR, NOR, SWE, USA, COT, CQU
M72056	4	FIN, NOR, RMO, SWE
M72068	3	FIN, NOR, SWE
M72076	9	FIN, FRA, MYS, NOR, QAT, RMO, RUS, TUR, COT
M72098	9	HKG, ISR, LTU, RUS, SGP, SWE, ADU, ARE, CQU
M72103	4	FIN, AAD, ARE, USA
M72121	10	FRA, ISR, RMO, RUS, SWE, ADU, AAD, ARE, USA, COT
M72170	2	GEO, PRT
M72180	4	GEO, HUN, ITA, LTU
M72190	3	RMO, ARE, CQU
M72198	4	LTU, MYS, NOR, SWE
M72209	3	TWN, FRA, SWE
M72227	7	FIN, HKG, HUN, LTU, MYS, NOR, CQU

Note: 'Count' indicates the number of countries in which the item is flagged as DIF item.

Table A.5: Results from the classical DIF study 5 on the responses of the PSI scale items.

Item ID	Description	Domain	Count	Countries
MQ12B01	Building Size	Applying	3	MYS[2], NOR, SWE
MQ12B02	Roof	Applying	7	ADU, ARE, HKG, ITA, MYS, SGP, TWN
MQ12B03	Constructing the Walls	Reasoning	6	CHL, FIN, MYS2, NOR, SGP, SWE[3]
MQ12B04A	Painting the Walls - area of side wall	Applying	6	ADU, HKG, HUN, RMO, RUS, TWN
MQ12B04B	Painting the Walls - total area	Reasoning	8	CQU, FIN, HKG, HUN, KOR, NOR, SWE, TWN
MQ12B04C	Painting the Walls - cost of paint	Reasoning	4	HKG, NOR, SWE, TWN
MQ12B05A	Water Tank - volume of tank	Applying	7	ENG[1], FIN, ISR, KOR, SGP, SWE, TUR
MQ12B05B	Water Tank - greater volume of tank	Reasoning	4	FIN, PRT, SGP, TUR
MQ12B05C	Water Tank - volume of cylinder	Reasoning	3	GEO, ITA, TWN
MQ12R01	Robots - rule	Reasoning	5	ADU, MYS, RUS, SGP, TUR
MQ12R02A	Robots - table for y	Applying	5	ITA, LTU, RUS, TUR[3], TWN
MQ12R02B	Robots - table for x	Applying	5	ADU, COT, ISR, TUR[3], TWN
MQ12R02C	Robots - different rule	Reasoning	7	ADU, HUN, ISR, KOR, MYS, TUR[3], TWN
MQ12D01	T-Rex Footprint Length to Foot Length	Applying	7	COT, ENG[1], NOR, PRT, SGP, TWN, USA[1]
MQ12D02AA	Relationship Between Foot Length and Leg Height - 37	Applying	11	FIN, GEO, HKG[2], ISR, KOR, NOR, RMO, RUS, SWE, TUR, TWN
MQ12D02AB	Relationship Between Foot Length and Leg Height - 61	Applying	8	FIN, GEO, ISR, KOR, RMO, RUS, TUR, TWN
MQ12D02AC	Relationship Between Foot Length and Leg Height - 78	Applying	8	FIN, GEO, ISR, KOR, RMO, RUS, TUR, TWN
MQ12D02B	Relationship Between Foot Length and Leg Height - formula	Reasoning	3	GEO, LTU, TUR
MQ12D03	Leg Height	Applying	12	ADU, FRA, GEO, HKG, HUN, ISR, KOR, RMO, RUS, SGP, TUR, TWN
MQ12D04A	Identifying Left and Right Footprints - bottom angle	Applying	2	QAT, RUS
MQ12D04B	Identifying Left and Right Footprints - foot	Applying	3	ARE, QAT, RMO
MQ12D05	Trexy's Stride Length	Applying	6	CHL, FIN, HUN, ISR, SWE, USA[1]
MQ12D06A	Ratio for a Dinosaur's Movement - shortest stride length	Reasoning	3	ARE, TUR[3], TWN
MQ12D06B	Ratio for a Dinosaur's Movement - longest stride length	Reasoning	4	GEO, KOR, TUR[3], TWN
MQ12D07A	Trexy's Running Speed - running speed	Applying	3	ADU, HKG, KOR
MQ12D07B	Trexy's Running Speed - faster dinosaur	Reasoning	0	
MQ12D08A	How fast could Trexy run a 100 m race? - 40 meters	Applying	5	COT, ENG, LTU, SWE, TWN
MQ12D08B	How fast could Trexy run a 100 m race? - entire race	Reasoning	4	GEO, HKG[2], SGP, TWN

Note: 'Count' indicates the number of countries in which the item is flagged as DIF item.

1: Effect further explored in Study 6 (ENG – USA); 2: Effect further explored in Study 6 (HKG-MYS); 3: Effect further explored (SWE-TUR) in Study 6

Table A.6: Number of countries with positive and negative DIF for eTIMSS and technology enhanced items.

	Regular items	**Required TEI**	**TEI as aid**
N-items	94	8	4
Mean number of countries positive DIF per item	1.90 (0-13)	0.5 (0-1)	1.25 (0-3)
Mean number of countries negative DIF per item	4.16 (0-16)	3.75 (1-8)	2.25 (1-4)

List of Figures

List of Tables