Validity evidence and measurement properties in technology-enhanced items

Part A: Building a validity argument

Authors

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

Contents

1. Abstract

Within this study we have addressed the question how to evaluate the validity of results of international large-scale assessment programs (ILSAs) that incorporate technology-enhanced items, with special attention for the comparability of results between countries. Within ILSAs the issues of validity and comparability are of utmost importance. These two cornerstones of methodology are closely connected to the two main goals of ILSAs: providing within-country trend comparisons and between-countries relative comparisons. The introduction of digital assessment in general and the use of technology-enhanced items more specifically offers the possibility to improve the authenticity and with that the validity of the measurement. Above that, technology-enhanced items could yield traces of (process) data that could be used to not only make statements about the proficiency of students, but also of the strategy that they have used in order to come to a response to a question. At the same time, the use of technology-enhanced items could have an impact on the comparability of country results and thereby jeopardizing the between-countries comparisons, the second main goal of ILSAs.

The study includes an interpretation and use argument to guide validation studies that are necessary to draw conclusions about the use of technology-enhanced items in TIMSS 2019. The validation studies include both qualitative and quantitative studies that aim to gather validity evidence.

It is concluded that the technology-enhanced items do not differ psychometrically from other digital items and that no additional differential item functioning (DIF) occurs. However, the qualitative studies show that the possibilities to achieve a better measure of problem solving are not met yet. The study ends with the conclusion that the current technology-enhanced items are still elementary and therefore, with these items, it is not possible to draw conclusions about the validity of advanced technology-enhanced assessments.

For ease of reading, the report was split into two parts. The interpretation and use

argument and the validity argument are described in this report, part A. The validity evidence is reported in detail in seven separate studies in part B of the report.

Keywords: Technology-enhanced items, validity argument, measurement variance.

2. Introduction

2.1 Evolution of international large-scale assessments

Over the past decade the use of international large-scale assessments (ILSAs) has shifted from their earlier use to measure learning at a national level for better understanding of the educational system towards a motivation for policy makers to improve their educational system. In some countries the results of ILSAs are used to give weight to arguments that are made to imply the need for drastic educational reforms (Lockheed and Wagemaker, 2013). Because of this growing importance of ILSAs, it is crucial that the validity of the results of ILSAs is guaranteed.

Following international innovations in assessments, ILSA organizations are moving towards digital assessments. IEA (International Association for the Evaluation of Educational Achievement), for example, states that 'Transitioning to digital assessment is important to "keep up with the times" and to increase both construct representation and data utility' (Cotter et al., 2020, p. 2.). One of the main surveys that is conducted by IEA is the Trends in International Mathematics and Science Study or TIMSS study. TIMSS monitors the trends in mathematics and science in more than 70 countries and benchmark participants[1]. Because not all TIMSS countries were prepared to conduct digital assessments, IEA decided to implement the transition over two assessment cycles—TIMSS 2019 and TIMSS 2023. In 2019, 32 of the 64 countries started administering TIMSS digitally[2] .

Now that TIMSS has gone digital, possibilities of computer-based testing can be fully used. For example, the 2019 cycle included both classical multiple choice and open-ended items, but also used digital item formats such as drag-drop and ordering items. More interesting is that several technology-enhanced items (TEIs) were included, in the form of extended Problem Solving and Inquiry tasks. These tasks aim to simulate authentic

[1] https://timssandpirls.bc.edu/
[2] https://timss2019.org/reports/about/

situations where several skills need to be integrated to solve mathematical problems (Cotter et al., 2020).

Psychometrically, TEIs hold the promise of improved measurement by being able to assess in more detail the exact strategy that students use to solve a given problem (Wools et al., 2019). This focus on different parts of the solution process in large-scale assessment contexts is impossible using paper-and-pencil items. The key of the promise of TEIs is that data is available not only about a student's response on an item of interest, but also about the different steps a student took to arrive at an answer. Specifically for mathematics, this yields possibilities because for classroom practices the strategies that students employ hold great didactic information.

The potential advantages of TEIs as part of educational assessment in general and TIMSS in particular are clear. However, there are potential threats to validity and measurement quality as well. One of the most prominent threats is mode effects. Is the same construct being measured when it is assessed on paper versus measured with a digital assessment that includes TEIs? To address this issue and to make sure the main outcomes of TEI based assessments were not affected by differences between paper and computer-based administrations, an item equivalence study and a bridge study were carried out (Fishbein et al., 2018; von Davier et al., n.d.). These studies focus on identifying and adjusting for potential differences between the two modes. However, additional validity questions can be raised with regards to TEIs. Especially in respect to mathematics, one might argue that working in an unfamiliar digital platform on complex math problems might affect the ability of a student to demonstrate their proficiency, and might therefore be a threat to validity. Students need to be able to work with tools like rulers and calculators to draw graphs or write formulas in a non-obtrusive fashion. Since this is not always the case, the transition from paper-based learning materials or the transfer from digital learning systems to the assessment system should raise validity questions that are to be addressed when interpreting results from large scale digital assessments.

For ILSAs an additional challenge arises. The results are not only used to make claims about student performance. The main purpose of these programs is to compare the performance of students from different countries. Little is known about the effect of TEIs with regard to the comparability for different groups of students. Since international comparable results are of utmost importance for ILSAs, it must be made clear as part of validation efforts that TEIs do not negatively impact comparability. Or even better, it must be made clear how TEIs can improve comparability or provide more insights in how countries compare.

2.2 This study

In this project we use the argument-based approach to validation (Kane, 2013) to investigate the impact of the use of technology-enhanced items in international large scale assessments. To this end we focus on three key aspects: the item content and assessment design, the psychometric properties, and identifying relevant population properties. In this empirical study the use of TEIs in TIMSS 2019 mathematics is evaluated by a multidisciplinary team. The results and methods used by the team are subsequently combined into guidelines for implementation and evaluation of TEIs in general and TIMSS mathematics

in particular. These guidelines will specify what type of technology-enhanced items as used in TIMSS mathematics function psychometrically sound under which conditions. The central research question of this study is: How can we evaluate the validity of the results of international large-scale assessment programs that incorporate technology-enhanced items, with special attention paid to the comparability of results between countries?

3. Theoretical framework

3.1 Validity and validation: the argument-based approach

In educational measurement, validity can be defined as the extent to which a test score is appropriate for the intended interpretation and use of the test (Kane, 2013). Within this definition a distinction is made between what a score means (the interpretation) and what is reported based on the test scores (the use). For test scores to be useful, the meaning that is added to a score goes beyond the test situation. For example, we would like to draw conclusions on a student's proficiency level of mathematics based on 30 items in a particular test form. Or, we would draw conclusions on a country's relative position of its performance in mathematics, based on a selection of items that was administered during a cycle of TIMSS. Whether it is warranted to draw these conclusions based on the actual observations that were made within the test administration is a matter of validity of test scores.

To evaluate the validity of test scores, Kane proposed the argument-based approach to validation (Kane and Brennan, 2006; Kane, 2013). The argument-based approach to validation aims to organize and structure validation efforts in a way that prioritizes the elements that threaten our ability to use and interpret the score as was intended (Wools et al., 2010). The main advantage of this approach is that in complex situations that involve assessments, research efforts are guided towards what is most appropriate instead of towards what is most common to evaluate.

To be able to decide what elements are most at risk in terms of validity of test scores, the argument-based approach to validation includes predefined stages. First of all, an interpretation and use argument (IUA) is drafted by the researchers. Within this IUA the intended interpretation and use of the assessment is made explicit in general terms. Subsequently, the IUA specifies the intended interpretation and use of assessment scores in greater detail. As part of this stage, underlying claims and assumptions are made explicit

and potential evidence to evaluate these claims is identified. These claims and assumptions are structured according to predefined 'inferences' (Wools et al., 2016).

The inferences that are specified within the IUA are flexible in a sense that particular uses of assessment results may call for different inferences. However, in general the following inferences are included in the IUA of educational assessments (Figure 3.1). These inferences aim to make the proposed interpretation and use of test scores more explicit by clarifying the underlying reasoning when we, for example, interpret performances on test forms as indication of the proficiency in mathematics of a student.

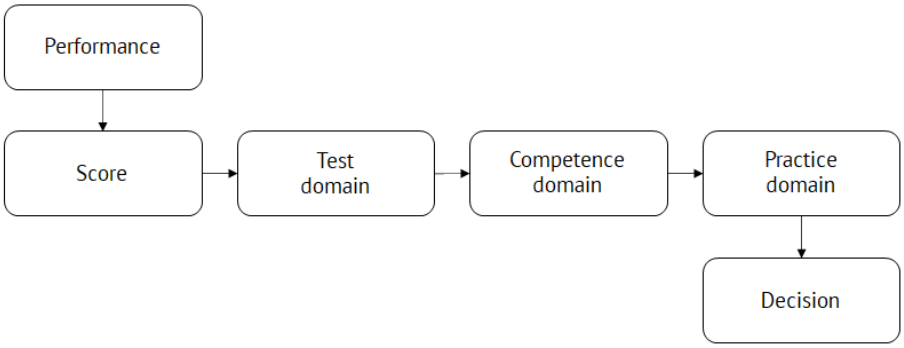

Figure 3.1: Baseline IUA: scoring, generalization, extrapolation I, extrapolation II, decision making (Wools et al., 2010).

The first inference in the baseline IUA relates to a performance on a task that translates into a numerical score. Subsequently, we infer that the score that is obtained on a particular test form can be generalized to a situation where other items could have been presented. The hypothetical collection of all possible items that could have been presented is referred to as a test domain. Once we interpret a score on a test form as a score on a test domain, the interpretation of the score is broadened even further. We infer that the test score can be extrapolated to a score on a competence domain, which entails an operationalization of the competence that is being measured. It is assumed that this competence domain is derived from a competence, ability or skill that is recognized and used in a real-life situation: a practice domain. Finally, this extrapolated score serves as input for a decision that is to be made about the competence of interest.

It is important to note that each inference can be seen as a practical argument in which the claim that is made in the preceding inference serves as the starting point for the next inference. All inferences are therefore linked together into one cohesive argument. Underneath every inference, assumptions and claims can be specified to help understand why certain inferences can be made. Following Toulmin's model for arguments (Figure 3.2), these assumptions and claims take the form of warrants, backings and rebuttals (Toulmin, 1953; Toulmin, 2003). Also, Toulmin's model for arguments provides us with opportunities to evaluate the assumptions and claims, and draw conclusions about the plausibility of inferences.

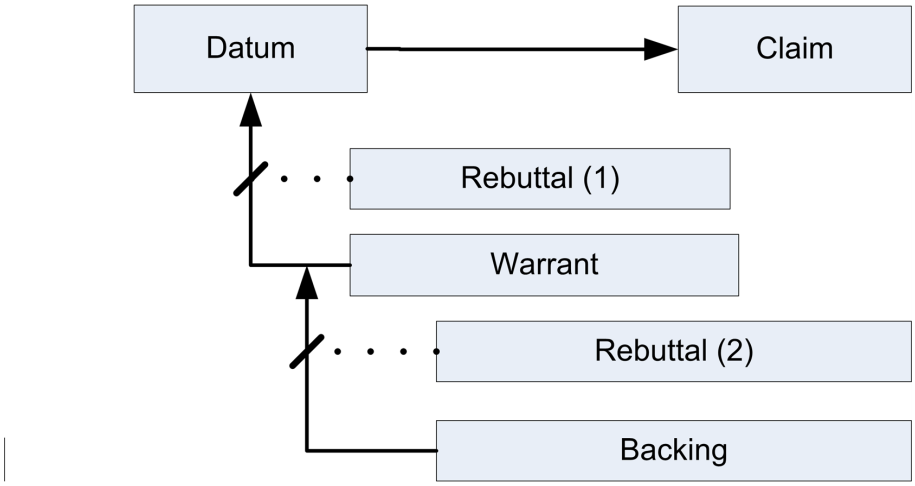

Figure 3.2: Toulmin's model for arguments.

Once an IUA is constructed, validity studies are designed and performed. To scope these studies one can use the inferences that are described according to Toulmin's model for argument. It makes sense to focus on the most questionable aspects, or aspects that are of major importance within the IUA first. When the validity studies are performed, the results and conclusions serve as input for the second stage of the argument-based approach to validation. In the second stage of the argument-based approach to validation a validity argument is constructed. This argument summarizes the available evidence regarding the validity of assessment scores and describes which claims from the IUA are rejected, accepted and for which claims additional research is required to draw a definite conclusion. To do so, a critical review of the available evidence is performed and conclusions are drawn regarding warrants, backings and rebuttals by the researchers of this study. This leads to a structured conclusion about the validity of test scores and their appropriateness for the intended interpretation and use.

3.2 International large-scale assessment programs

International large-scale assessments (ILSAs) of education aim to inform policymakers, educational researchers, and the general public. In general, ILSAs are empirical studies in which student achievement is assessed and contextual information about school systems is collected (Hastedt and Rocher, 2020). The results of the achievement tests and the contextual questionnaires are combined to draw conclusions about educational systems. To put these conclusions in perspective, results are presented in a way that enables comparisons between countries. Several impact studies have shown that ILSA results have been used to support policy making (e.g. Breakspear, 2012; Schwippert and Lenkeit, 2012;Wagemaker, 2013). Hastedt and Rocher (2020) describe that various educational improvements have been supported by evidence from ILSAs that are reported in the TIMSS and PIRLS Encyclopedias. The data from ILSAs provide valuable opportunities to help inform both policy decisions and research into education system improvement. The most well-known organizations that develop ILSAs are the OECD (Organization for Economic Co-operation and Development), and IEA (International Association for the Evaluation of Educational Achievement) (Hastedt and Rocher, 2020). The two organizations have

different approaches, in terms of study philosophy, content selection and cohort selection. With regard to the study philosophy IEA studies "seeks to measure what is taught in schools and the contexts of learning" (Hastedt and Rocher, 2020, p. 3), while OECD PISA "seeks to measure selected acquired skills of studies towards the end of their compulsory education" (Hastedt and Rocher, 2020, p. 3). In line with the different philosophies the content selection for IEA is based on the curricula of participating countries, while for OECD PISA the content is selected by experts. Therefore the conclusions that can be drawn for each of the studies differs. In this project the focus is on the TIMSS study from IEA.

3.3 Innovations in ILSAs

All over the world, the use of technology in education is increasing significantly. These technological advancements do not only benefit learning materials, also assessment practices are innovated (Wools et al., 2019). And thus, following international innovations in assessments, international large-scale assessment organizations are moving towards digital assessment as well. As ILSAs continue to modernize, new methodological opportunities and challenges lay ahead (Hastedt and Rocher, 2020).

In 2019, TIMSS was conducted in part online with e-TIMSS test forms (Mullis, 2017). Digitizing assessments can improve them by using more authentic items that have the potential to improve construct representation, making it possible to assess complex constructs like skills or competences (Sireci and Zenisky, 2011). This aligns with the focus of ILSAs to explore the possibility to assess so called '21st century skills' (Hastedt and Rocher, 2020). Authentic items have become increasingly more complex and are often referred to as Technology-Enhanced Items (TEIs), or as defined by the Measured Progress and ETS Collective (MeasuredProgress and ETS, 2012, p. 1): *"Technology-enhanced items (TEI) are computer-delivered items that include specialized interactions for collecting response data. These include interactions and responses beyond traditional selected-response or constructed-response."*

With TEIs, meaningful additional data is gathered beyond the traditional correct or incorrect response. This additional data includes for example log-files, time stamps and chat histories (Wools et al., 2019). Although gathering new data could be helpful in gaining better insights of students abilities, this would only be the case when this data could be interpreted validly. This provides us with challenges and research is necessary regarding the topic of "process data" (Hastedt and Rocher, 2020).

3.4 Differential item functioning

If the item characteristics of ILSA items differ across countries, one speaks of differential item functioning (DIF) or measurement variance (Thissen et al., 1993). In general, DIF is undesirable as -in the presence of DIF- it is psychometrically difficult to interpret differences between countries in the knowledge domain represented by the items. That is, for a meaningful comparison, the item characteristics (as operationalized by psychometric properties like item difficulty and item discrimination) need to be the same in the countries of comparison, which would mean the absence of DIF. In practice, DIF can both be seen as an indication of item bias (e.g. Mellenbergh, 1989) 1) due to for example a country having

an unfair advantage on an item that has no clear interpretation in terms of the actual knowledge domain being assessed by the items, or, 2) DIF can be seen as a valuable source of information concerning differences across countries (Verhelst, 2012; Zwitser et al., 2017). For instance, in case of the latter, DIF may indicate differences in educational policy across countries, which is arguably one of the most important outcomes of an ILSA, or DIF can indicate differences in the solution process across countries. When we want to assess the validity of educational tests in general, in particular the validity of incorporating technology-enhanced items in ILSAs, DIF is of key importance for ensuring that a test is sensitive to relevant country differences (e.g. education policy differences, and differences in response process) and insensitive to unimportant differences (i.e., confounding or biasing effects). The choices related to DIF during the scaling process can have a substantial impact on the outcomes of ILSAs (Feskens et al., 2019; Robitzsch and Lüdtke, 2020; Jerrim et al., 2018). Therefore, in the present project, the assessment of DIF plays a prominent role in evaluating the validity of ILSAs with technology-enhanced items. In general, we focus on three aspects related to DIF. That is, we assess: 1) If the item responses demonstrate DIF across countries; 2) If the on-screen times (as a source of process data that give an indication of the response process or solution strategy needed to solve the item) demonstrate DIF across countries; and 3) If the DIF in the item responses across countries can be explained by differences in the on-screen time.

4. Methods

In this project we use an argument-based approach to validation to investigate the impact of the use of technology-enhanced items in international large-scale assessments. To do so, a validation study is conducted that will eventually lead to recommendations regarding the use of technology-enhanced items in ILSAs. The validation study follows the argument-based approach to validation by Kane (2006; 2013). In this approach three stages can be identified that are linked to three research activities:

1. Develop an interpretation and use argument
2. Gather validity evidence
3. Combine evidence into a validity argument

For readability purposes the report was split into two parts. The interpretation and use argument, the validity argument and the conclusions are described in this report, part A. The validity evidence is reported in detail in seven separate studies in part B of the report. The conclusions from these studies are used in the validity argument, as described in this part of the report. The ILSA that will be used 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 scope of this study in more detail. Subsequently, we discuss the three research activities.

4.1 Scope of this study

4.1.1 Assessment of interest: TIMSS 2019 - Mathematics grade 8

This study focuses on performing a validation study of the mathematics part of TIMSS 2019, specifically grade 8. 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 provides data users with information about how achievement compares among countries and whether scores are improving or declining over time (Martin et al., 2020).

4.1.2 eTIMSS

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.

4.1.3 Problem solving and Inquiry tasks (PSI)

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).

4.2 Developing an interpretation and use argument

4.2.1 Constructing an interpretation and use argument

When constructing an interpretation and use argument (IUA), information about an assessment is gathered and structured according to predefined inferences. In this report, a document study was conducted to get an insight into the design, rationale and results of TIMSS 2019. The IUA was constructed based on the information available to the authors of this report and within the context of this report.

The following documents were consulted:

- TIMSS 2019 Assessment Framework – Chapter 1 TIMSS 2019 Mathematics Framework (Lindquist et al., 2017);
- TIMSS 2019 International Results in Mathematics and Science (Mullis et al., 2020;
- Findings from the TIMSS 2019 Problem Solving and Inquiry Tasks (Mullis et al., 2021);
- Methods and Procedures: TIMSS 2019 Technical report (Martin et al., 2020).

For every inference in the validity argument, relevant information was clustered and selected.

4.2.2 Identifying claims & assumptions at risk

When the argument was designed in general terms, claims were formulated according to the available information. Each claim was then labeled as 'warrant', 'backing', or 'rebuttal'. In some cases, multiple warrants and backings were identified within one inference. Wools et al. (2010, p. 66-67) describe that the basis of an argument is the distinction between the claim we want to establish and the facts that serve as the foundation of the claim. Once the facts are provided, it may not be necessary to provide more facts that can serve the claim. Moreover, it is important to state how the facts lead to the claim that is being made. The question to be asked should not be 'what have you got to go on?', but 'how do you get there?'. Providing more facts of the same kind as the initial facts is not appropriate to answer the latter question. Therefore, propositions of a different kind should be raised: rules or principles. By means of these rules or principles, it can be shown that the step from original data to the claim is legitimate. The rules and principles will thus function as a bridge from data to claim. These bridges are referred to as warrants. As the warrants possess neither authority nor currency, the distinction between data, on the one hand, and warrants, on the other, is not an absolute distinction since some warrants can be questioned. Supporting warrants are assurances referred to as backing. Lastly, Toulmin (Toulmin, 1953; Toulmin, 2003) mentions a rebuttal, which indicates circumstances in which the general authority of the warrant would have to be set aside. A rebuttal provides conditions of exception for the argument.

One aspect in developing an interpretation and use argument is the identification of the inferences that are most at risk. Validation is sometimes perceived as an endless endeavor. The argument-based approach to validation aims to scope the validation efforts by prioritizing the claims that need evidencing the most. The present study focuses on the implementation of technology-enhanced assessment. This focus was taken into account when selecting the claims that were scrutinized in this study.

4.3 Gathering validity evidence

After identifying the claims and assumptions at risk, seven studies were performed to gather additional validity evidence. The studies were designed so that they would provide evidence for the claims made by the inferences of the interpretation and use argument. For ease of reading, each of these studies, including their methods, is described in part B of this report.

4.4 Combining evidence into a validity argument

The final stage of developing an interpretation and use argument involves a critical appraisal of the available evidence. The results of the seven studies were used to draw conclusions about the claims in the interpretation and use argument. Are warrants, backings and rebuttals accepted or rejected, or is additional research required to draw a conclusion? The answers to these questions were incorporated into the inferences to show validity flaws or strengths. For some inferences visual representations were made.

5. Results - Interpretation and use argument

5.1 Inferences within interpretation and use argument for TIMSS 2019

This chapter presents an interpretation and use argument (IUA) for the mathematics part of TIMSS 2019 with specific attention to the innovative technology-enhanced items that were included in this cycle. The argument describes the intended interpretation and use of the assessment scores of TIMSS. An IUA consists of several connected inferences that help understand how one can reason from the performance of a single student on a test form to decisions about the mathematics proficiency of a student population of a country as a whole. In general, the interpretation of TIMSS could be summarized as a reflection of math proficiency as taught in schools and the context of learning. The conclusions from TIMSS aim to inform policy makers by providing robust information that is largely independent of any single political system (Hastedt and Rocher, 2020). To help the public understand and contextualize the results, results of TIMSS are often presented in a rank order table, making comparisons between countries possible.

The inferences that are included in the IUA of TIMSS 2019 differ somewhat from the baseline IUA presented in Figure 3.1. This is because IEA specifies quite explicitly that the aim of TIMSS is to make claims about "math proficiency as taught in schools and the context of learning". By stating this, the scope of TIMSS is explicitly limited to a competence domain and not a practice domain. This leaves us with a scoring inference (Performance - Score), a generalization inference (Score - Test domain), one extrapolation inference (Test domain - Competence domain) and a decision inference (Competence domain - decision).

The remainder of the IUA consists of an in-depth description of inferences, claims and assumptions underlying the general interpretation and supporting the intended use. Since these inferences, claims and assumptions differ depending on the administration modes (or assessment types), a distinction is made for paperTIMSS, eTIMSS, and PSI. This distinction complicates the IUA in a way that the inferences could be unique for

each assessment type. Therefore, the general structure of the IUA takes the form of an assessment program (Wools et al., 2016) where an original interpretive argument reasons from one performance to a decision. The extended interpretive argument for assessment programs can, however, incorporate multiple performances, multiple test domains, and multiple competence domains that are aggregated into one decision.

For TIMSS, in which three assessment types are distinguished, the shape of the interpretation and use argument for the three assessment types is visualized in Figure 5.1.

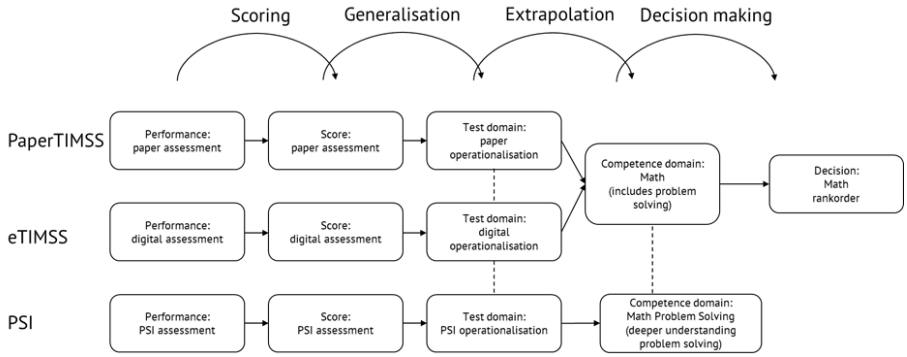

Figure 5.1: Inferences within the interpretation and use argument TIMSS 2019.

In Figure 5.1 three types of performance are specified: students taking the assessment on paper (paperTIMSS), students taking the assessment in a digital environment (eTIMSS) and students taking the innovative PSI booklets in a digital environment. For all delivery modes, performances are translated into a score (scoring inference). This score is generalized into a score that is representative for all possible tasks that are included in the test domain (generalization inference). Since there is considerable overlap between the item bank of paperTIMSS and eTIMSS one could argue that the test domain for those two conditions is identical. However, during the development of eTIMSS an effort was made to develop some items that are supported by the digital medium (Martin et al., 2020). Therefore, it was decided to work with two different test domains for paperTIMSS and eTIMSS. The PSI condition is also treated separately since these items were constructed with additional guidelines. Also, limitations of the regular eTIMSS environment that stem from comparability were loosened. This allowed for items that aim to extract not only answers but behavior could be used to make inferences about strategy use as well. When the scores on the test domain are extrapolated to the competence domain, the scores of paperTIMSS and eTIMSS are merged into one competence domain: a score on the described construct of math that includes problem solving as a sub-competence. The consequence of this merge is that assumptions are made about the comparability of paperTIMSS and eTIMSS. These assumptions are especially strong when it comes to comparability of scores for the test domains and competence domains. Figure 5.1 visualizes these assumptions through the dotted lines.

PSI tasks intend to measure a deeper level of problem solving and the tasks include a broader array of innovative digital features than regular (e)TIMSS items (Martin et al., 2020). Thus, even though the PSI tasks and the (e)TIMSS items are part of the same framework, the intention is to extent the measurement of problem solving with PSI tasks and to use different item types. This difference is expressed by recognizing two different

competence domains. This difference is also accounted for in the decision inference. For TIMSS 2019 the results on the PSI items were presented separately from the regular eTIMSS scores and results on these items were not included in the scale scores used for trend analysis as initially presented by TIMSS 2019 (Martin et al., 2020). In an additional report, the PSI scores were presented as part of the scale scores (Mullis et al., 2021). In the IUA this is reflected by a decision inference that only includes scores from paperTIMSS and eTIMSS. Since there were also no decisions made about a student's or country's level of proficiency on PSI tasks alone, a separate decision inference was not required for the PSI tasks.

5.2 Identifying claims

Now that the inferences are specified, the claims underlying the inferences can be identified. These claims aim to support the inferences and are subject to validation studies. The results of the validation studies will be used to decide whether claims are to be accepted, rejected or that additional evidence is still required. In this particular study, we aim to investigate the impact of technology-enhanced items on the validity of TIMSS. Therefore, the claims that are subjected to research are all related to the use of these innovative items. The validity of TIMSS 2019 as a whole is beyond the scope of this study. An example of studies on the validity of TIMSS can be found in Wagemaker (2020).

All inferences are structured according to Toulmin's model for arguments. Therefore, within an argument a distinction is made between datum and claim, warrant, backing, and rebuttal. In this study, the focus of the inferences was defined in line with the earlier described rationale and scope of this study: technology-enhanced assessment. As technology-enhanced items are only apparent in the PSI and the eTIMSS argument, we focus solely on the claims for eTIMSS and PSI. However, as mentioned earlier, the aim of TIMSS is also to compare results between different administration modes. Therefore, we will also look into the comparison between eTIMSS and paperTIMSS and between eTIMSS and PSI. Figure 5.2 shows the inferences that will be addressed in this study in black, the ones that are considered out of scope are grey.

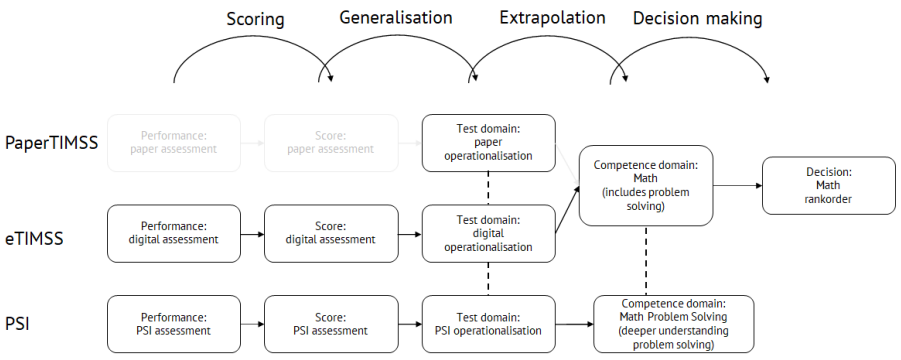

Figure 5.2: Inferences and claims included in this study.

5.3 Inferences eTIMSS

The inferences that were described for eTIMSS are shown in Figure 5.3. For every inference, at least one warrant and backing are specified and, when applicable, a rebuttal.

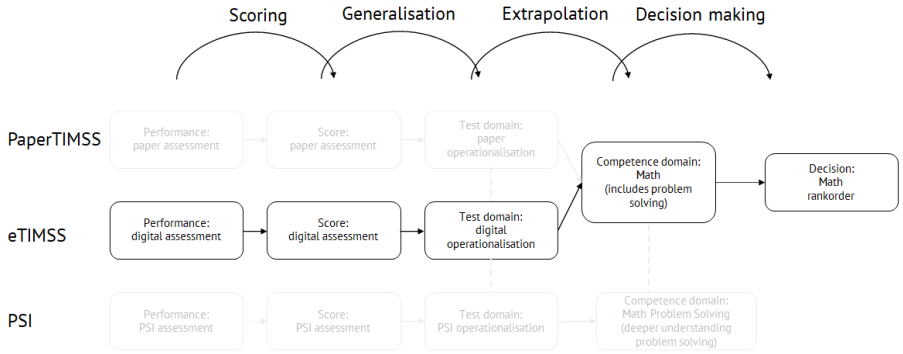

Figure 5.3: Inferences and claims regarding eTIMSS assessment.

5.3.1 Scoring inference eTIMSS

The scoring inference aims to explain how a score is derived from a student's performance. It is quite common for assessment programs in education to mark answers of students as correct or incorrect based on a marking scheme. This is also the case with TIMSS (Cotter et al., 2020). The marking scheme makes sure that similar performances lead to similar scores. In TIMSS this is especially relevant for the constructed response items. Since eTIMSS is a digital platform, it is also assumed that all answers are stored and scored correctly and that no technical difficulties that could cause a loss of data were registered. Figure 5.4 shows these claims in the Toulmin model.

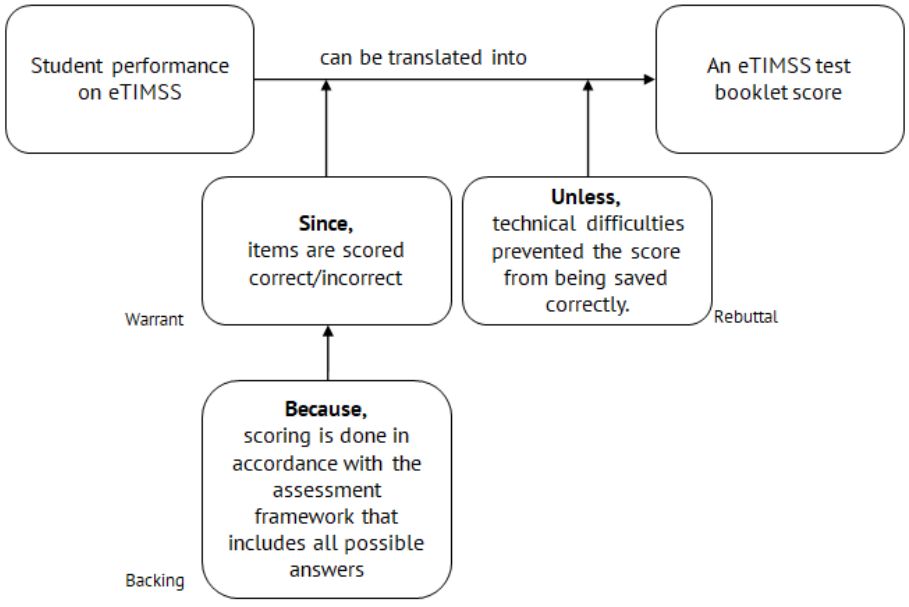

Figure 5.4: Scoring inference eTIMSS.

5.3.2 Generalisation inference eTIMSS

Within the generalisation inference it is argued that a score on one eTIMSS test form can be generalised into a hypothetical score on the test domain of math. A test booklet consists of a sample of items, the performance on which one would like to generalise to a larger domain of items that were not presented. But what would have been the performance of this student if they had been given another test booklet? And, it could also be the case that many items remain unanswered, and particular content elements are not included in students' performances. This would jeopardise the generalisation inference. Beside a content perspective, one could also take a more statistical perspective on generalisation. For example, one could state that the sample of items in a test booklet should at least be large enough to control for measurement error, as can be expressed in a reliability coefficient. Figures 5.5 and 5.6 show these rationales in the form of a Toulmin model.

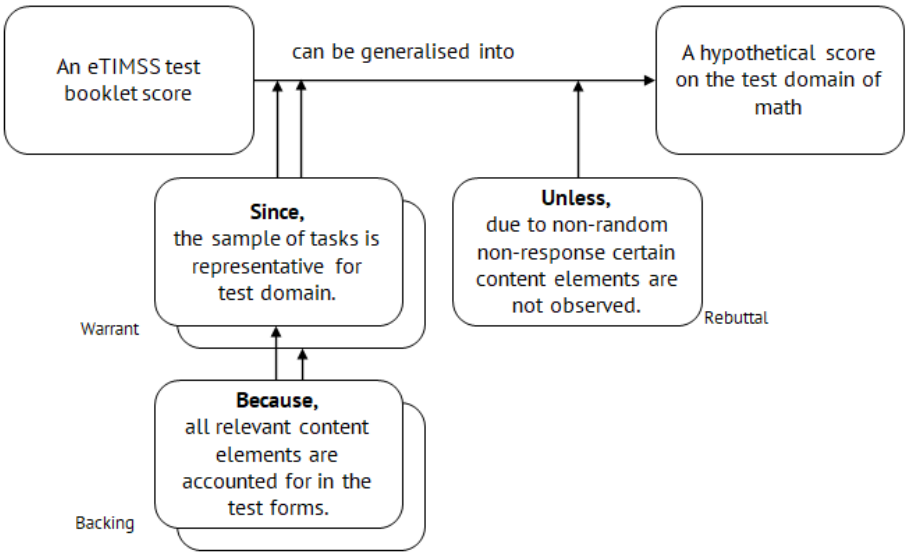

Figure 5.5: Generalisation inference eTIMSS 1/2.

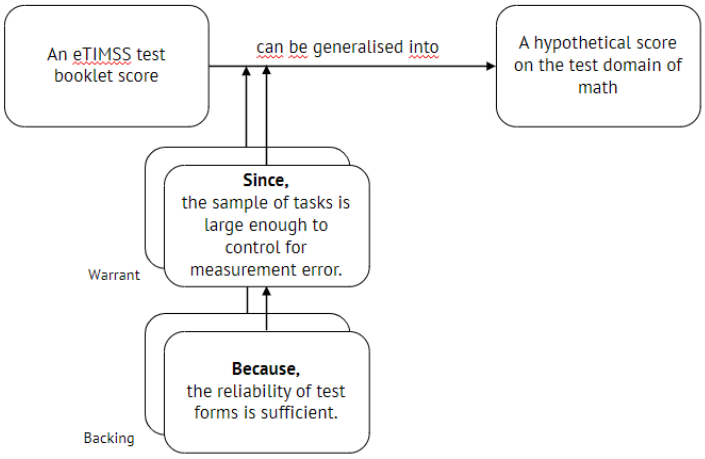

Figure 5.6: Generalisation inference eTIMSS 2/2.

5.3.3 Extrapolation inference eTIMSS

The extrapolation inference makes it explicit that conclusions about math can be drawn based on the items included in eTIMSS. This questions the operationalisation of math into the tasks that were included in eTIMSS. This is because we also assume that the items include the use of critical aspects of the competence of interest. However, it could be possible that other aspects besides the competence of interest cause variance in the scores, for example, digital literacy, reading proficiency or even cheating. These aspects cause construct irrelevant variance, something that will lessen the possibility to extrapolate the scores on the test domain to the competence domain. Next to the quality of the operationalisation, one can question the quality of the competence description. It could be possible that all items match a certain competence description that is in itself very narrow. TIMSS aims to measure what is taught in schools and the context of learning. The competence domain of math should therefore be recognized by experts as being the construct of math as taught in schools. These assumptions are visualised in a Toulmin model in Figures 5.7 and 5.8.

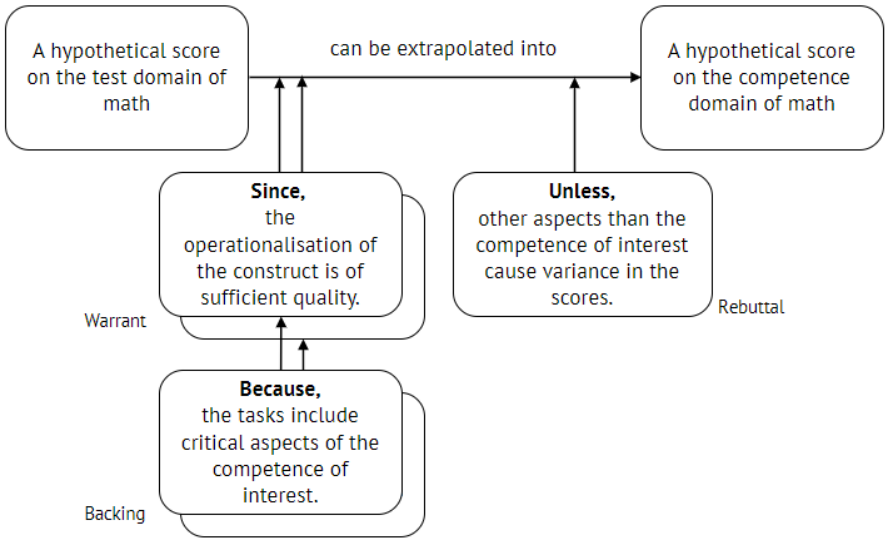

Figure 5.7: Extrapolation inference eTIMSS 1/2.

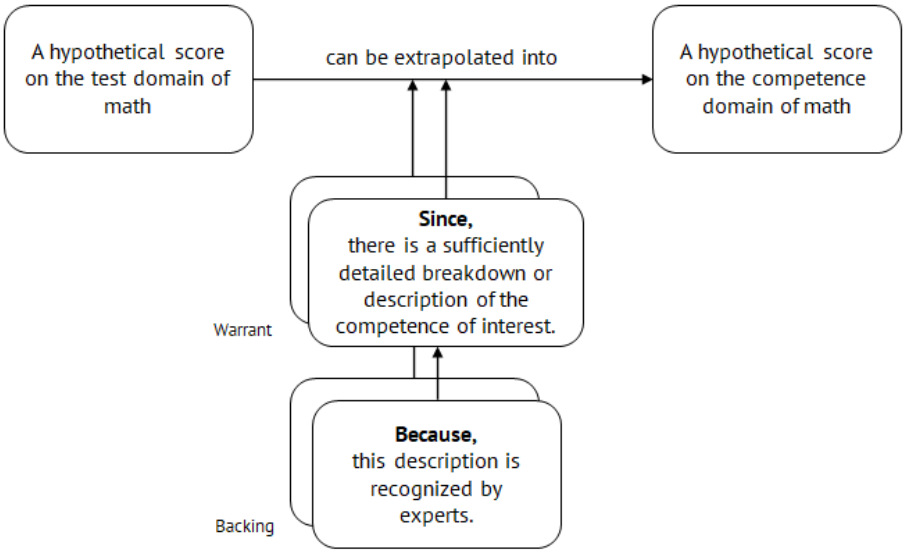

Figure 5.8: Extrapolation inference eTIMSS 2/2.

5.3.4 Decision inference eTIMSS

In the final inference, it is assumed that the hypothetical score on the competence domain can be used for decisions about trends in math education and curricula. This is because TIMSS uses scale scores that are comparable between countries and earlier cycles. Therefore, policy makers can evaluate their curriculum and the impact of policy decisions on learning outcomes. These assumptions are visualised in Figure 5.9.

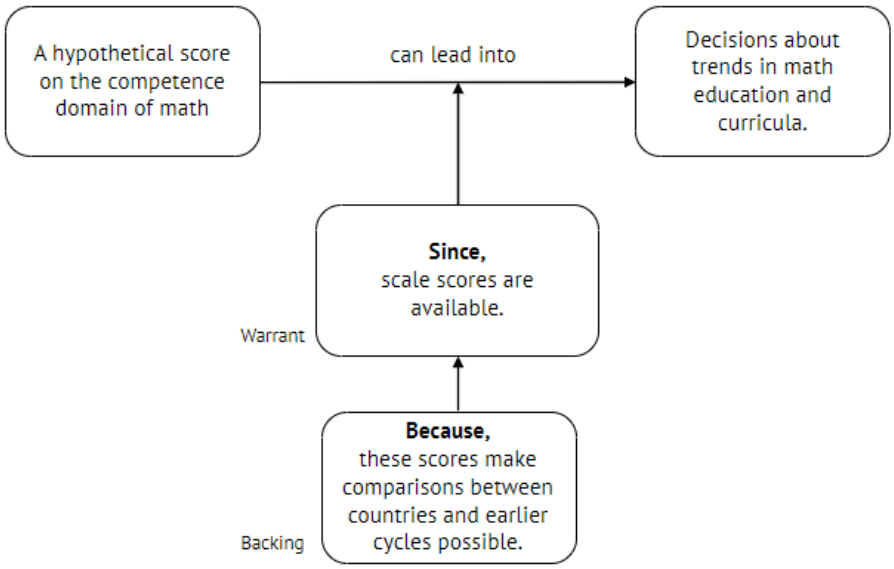

Figure 5.9: Decision inference eTIMSS.

5.4 Inferences PSI

The results of PSI booklets are not used to establish the trend results in TIMSS 2019 (Mullis et al., 2021). Therefore, the decision inference is not applicable for these booklets. Figure 5.10 shows the inferences that are applicable for PSI.

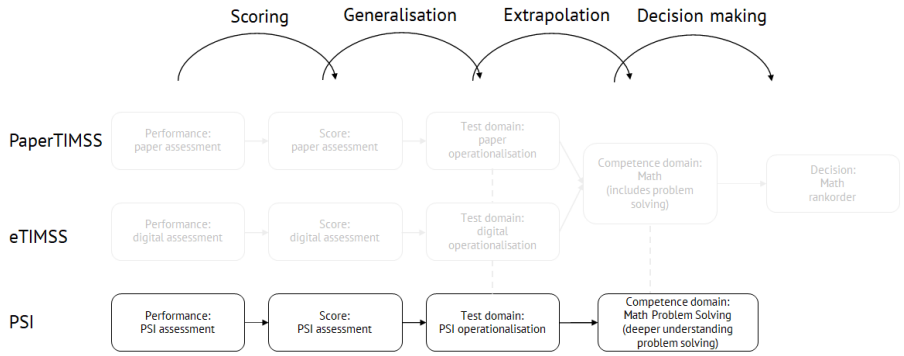

Figure 5.10: Inferences and claims regarding PSI assessment.

The assumptions underlying the inferences of PSI are the same as the assumptions within the inferences of eTIMSS. Although PSI is part of the measurement of mathematics, the main difference is that the intended competence for PSI is 'problem solving skills' instead of 'math'. This affects the competence domain and inferences regarding the competence domain. In the inferences of eTIMSS, the word 'math' should be replaced with the words 'problem solving'.

5.5 Inferences comparability

Since the performances are different, the three item types (paperTIMSS, eTIMSS, PSI) are distinguished in this interpretation and use argument. However, assumptions are also made

about the comparability of the results (test domain and competence domain). Therefore, assumptions underlying the comparability of these inferences are made explicit in this interpretation and use argument (Figure 5.11).

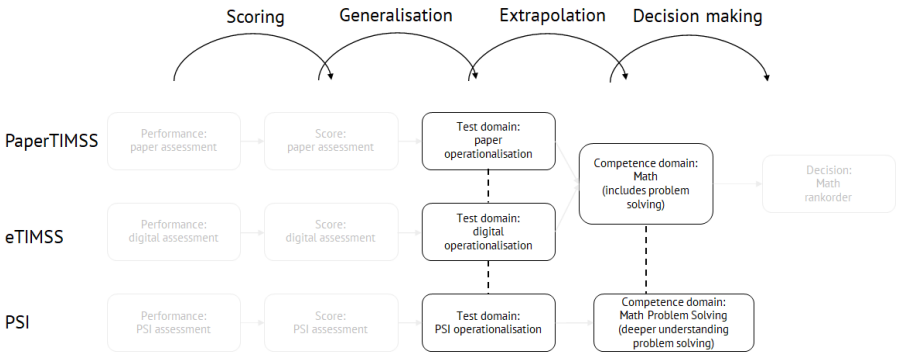

Figure 5.11: Inferences and claims regarding comparison between paperTIMSS, eTIMSS and PSI.

5.5.1 Comparability test domains

The inferences made to draw comparable conclusions about the different test domains involve claims about the quality of the operationalisation. In this particular IUA, we distinguished three test domains that describe (parts of) the same construct. The rationale behind this choice is that the assessment type (paper, computer, PSI) holds different possibilities to assess the construct of interest. Therefore, the operationalisation of the construct into tasks could differ for each administration mode. However, since we would like to compare the conclusions that are drawn based on this operationalisation, we assume that the only difference in the operationalisation is related to administration mode and constraints that occur for this administration mode. It is important to establish that from a content perspective, the operationalisation does not differ. Figures 5.12 and 5.13 show the inferences for the comparisons of the test domain.

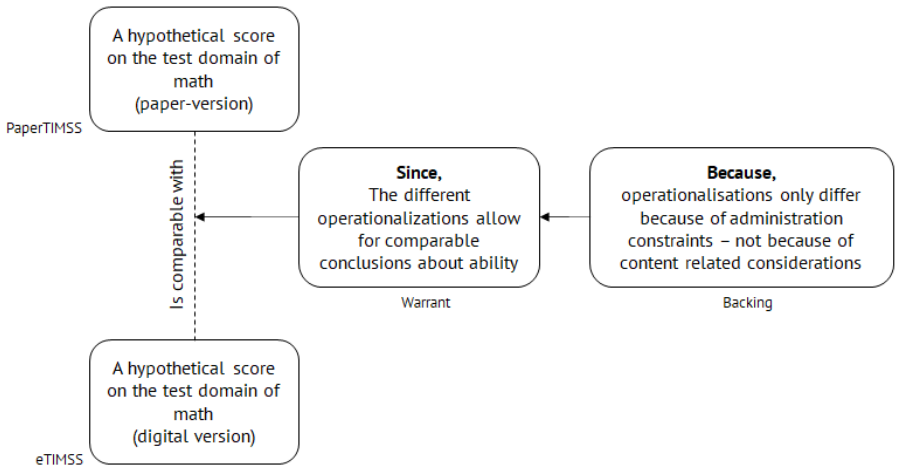

Figure 5.12: Inference comparison test domain paperTIMSS - eTIMSS.

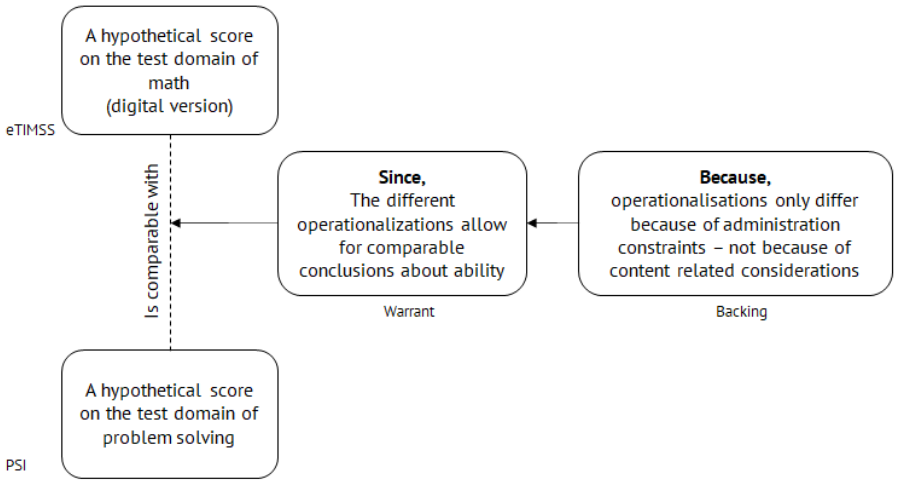

Figure 5.13: Inference comparison test domain eTIMSS - PSI.

5.5.2 Comparability competence domains

In Cotter et al. (2020) it is described that the PSI tasks aim to measure problem solving as a part of math competencies but at a deeper level than paperTIMSS and e-TIMSS. Also, PSI tasks do not aim to measure the full competence of math, as they are focused on the aspect of problem solving. Therefore, within the IUA two competence domains are distinguished: math and a subset of math, namely problem solving. It is assumed though, that both competence domains are comparable in a sense that the broader concept of math includes problem solving and that within the PSI tasks, problem solving is only isolated from other math subskills.

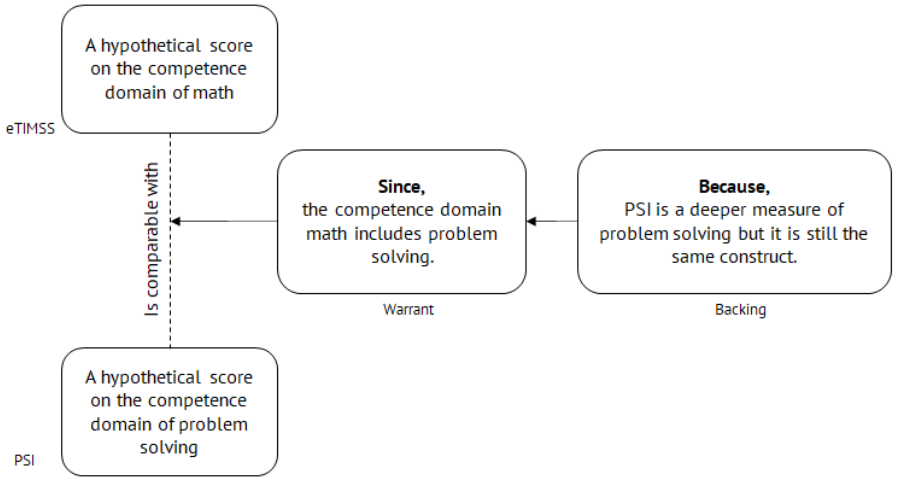

Figure 5.14: Inference comparison competence domain eTIMSS - PSI.

5.6 Identifying sources of evidence

In this stage, inferences are studied and the claims or assumptions that are most important or are questionable are identified. Subsequently, it was decided what evidence would be necessary to support or reject these claims. To identify evidence that was readily available

and well known, background documents were studied. Based on these findings, seven studies were designed to collect the required additional evidence. The studies and their methods are described in part B of the report.

6. Results - Validity argument

This chapter demonstrates how a validity argument can be constructed, based on the interpretation and use argument (IUA) as presented in Chapter 5. The aim of the performed studies that was to gather the evidence that was most needed in relation to the specific scope of this research project. These studies are separately presented and discussed in part B of the report. Some claims are evaluated based on evidence that is publicly available on the IEA website or in other publications.

All inferences underlying the conclusions are presented in Table 6.1, 6.2 and 6.3. However, note that in the validity argument a distinction is made between claims that are related to the focus of this study (the addition of technology-enhanced items, PSI tasks and comparability between countries) and claims that are not.

The claims that are related to TEI and PSI tasks were evaluated on the basis of existing evidence and the additional evidence gathered in the studies presented in part B of the report. On the basis of this evidence, it was decided to accept, reject, or withhold drawing a conclusion (i.e. indecisive) about the warrants, backings and rebuttals. This enabled us to draw conclusions about these arguments. A rejection of a warrant or backing means that the inference as a whole should be rejected. As for rebuttals, acceptance of a rebuttal would cause an inference to be rejected. In case it was not possible to gather the required evidence or if additional research was necessary, no definite conclusion could be drawn. In terms of the visual representation, when an inference or claim is accepted the arrow connecting the elements is solid; when it is rejected or indecisive the arrow is dotted.

Most claims regarding eTIMSS are not related to the scope of this study. These claims were broadly evaluated on the basis of publicly available evidence (found on the IEA website or in other publications, (e.g. Cotter et al. (2020)). Often, this led us to conclude that both the backing and warrant could be accepted and the rebuttal could be rejected. Or in other words, that the inference could be considered as supporting the validity of the intended interpretation and use of eTIMSS. In the case that no evidence was available, the

inference was marked as 'out of scope'.

6.1 Inferences for eTIMSS

Table 6.1 presents an overview of the claims, evidence and conclusions for eTIMSS. For the scoring, generalisation and the second extrapolation inferences of eTIMSS, the evidence required to draw a conclusion was available. Some of this evidence is summarized and described in Study 2 and 3. For these inferences the warrants and backings are accepted, and the rebuttals are rejected. The inferences as a whole are therefore valid. Additional evidence was gathered for two claims (marked in bold in Table 6.1), and will be discussed in more detail.

Table 6.1: Evidence for Validity Argument of eTIMSS.

Claim	Evidence	Study	Conclusion
	Scoring inference (Figure 5.4)		
Warrant	Documents to support the scoring method	Available: described in documentation on website	Accepted
Backing	Assessment framework that includes answers and scoring manual	Available: described in documentation on website	Accepted
Rebuttal	Administration logs	Out of scope	Indecisive
	Generalisation inference 1 (Figure 5.5)		
Warrant	Description of construct	Available, and described in Study 2	Accepted
Backing	Test matrix and test version design	Available, and described in Study 2	Accepted
Rebuttal	Non response analysis	Available, and described in Study 3	Rejected
	Generalisation inference 2 (Figure 5.6)		
Warrant	Number of items per content domain	Available, and described in Study 2	Accepted
Backing	Reliability analysis	Available, and described in Study 2	Accepted
	Extrapolation inference 1 (Figure 5.7)		
Warrant	Description of Math operationalisation	Available: described in documentation on website	Accepted
Backing	Description of Task design and item analysis	Available: described in documentation on website	Accepted
Rebuttal	**Analysis of potential sources of variance**	Available: described in documentation on website **Additional: Study 6**	**Rejected**
	Extrapolation inference 2 (Figure 5.8)		
Warrant	Description of the construct of interest (Math)	Available: described in documentation on website	Accepted
Backing	Description of procedure that involves expert judgements	Available: described in documentation on website	Accepted
	Decision inference (Figure 5.9)		
Warrant	Description of procedure to estimate scale scores	Available: described in documentation on website	Accepted
Backing	**Analysis regarding comparability**	Available: described in documentation on website **Additional: Study 5**	**Accepted**

For the rebuttal in the first extrapolation inference additional evidence was gathered, unless other aspects than the competence of interest cause variance in the scores. One

way of showing construct irrelevant variance is through DIF studies. Study 5 (Classical Differential Item Functioning) shows that there is not significant DIF between countries, leading us to conclude that this particular source of construct irrelevant variance is not present. Furthermore, on the basis of study 5 and 6, it was concluded that the difficulty and response times of these items were within the range of non-technology enhanced items. On the basis of this finding and the results from the bridge study (Hamhuis et al., 2018), we concluded that the rebuttal should be rejected. We conclude that the inference is valid, which is visualized in Figure 6.1.

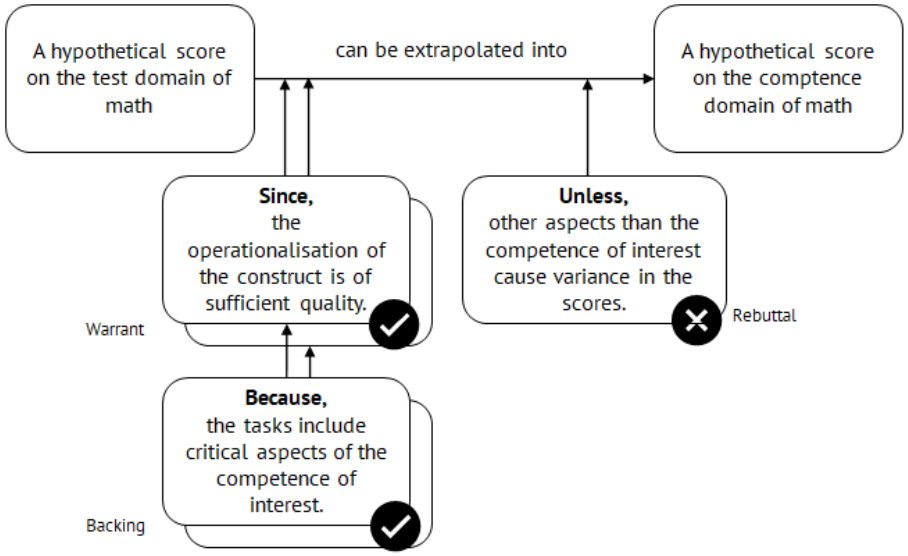

Figure 6.1: Validity Argument: eTIMSS extrapolation 1/2.

In the final inference of the eTIMSS validity argument, claims focus on comparable scores: scores that can be compared over time and over countries to enable decisions about international trends. The evidence that is considered here comes partly from the manual (availability of scale scores). However, the backing calls for comparison between countries and this particular claim was studied in more depth. Study 5 (Classical differential item functioning) and Study 6 (Process data) aimed to draw conclusions about the possibility to validly compare scores. Both studies did not show large differences for subgroups that would cause us to conclude that comparisons are not possible. On the scoring level as well as on the response times level, we do not see big differences between countries. Therefore we conclude that comparisons on the same construct are possible. This is visualised in Figure 6.2.

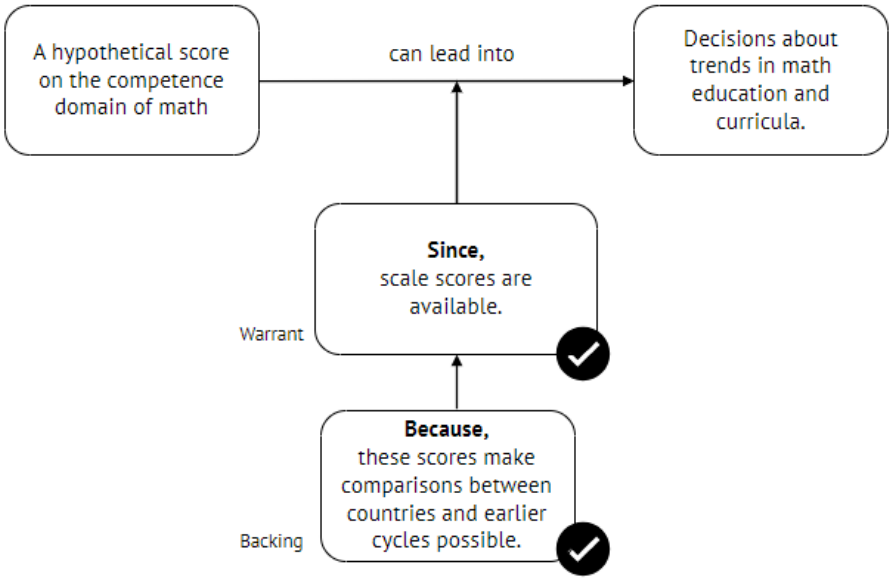

Figure 6.2: Validity Argument: eTIMSS decision.

6.2 Inferences for PSI

The claims within the IUA for PSI are very similar to the IUA of eTIMSS. However, the evidence underlying the claims is not identical. The inferences are addressed separately in this section. Also, conclusions differ from the conclusions drawn for eTIMSS because other procedures are used. Different design choices are made with different evidence as a consequence. An overview of the claims in PSI and the additional evidence is provided in Table 6.2.

6.2.1 Scoring inference PSI

The current procedures of scoring within PSI do not deviate from the usual eTIMSS scoring procedures. The final answers are scored correct, partially correct, or incorrect. This is done automatically for most items. Therefore, there is not much reason to doubt the validity of the scoring. Constructed item responses are scored by using the the IEA Coding Expert Software that incorporates the IEA standards (Johansone, 2020). To assure reliability, scorers are trained in using the scoring scheme. However, technology-enhanced items hold the promise that other scoring methods can be used to get a deeper understanding of the construct. That it, for example, would be possible to not only score correct/incorrect but also to identify the strategy that is used to reach an answer. In the PSI items this was not the case, therefore this could not be studied. We accept the current warrant and backing, and reject the rebuttal. We therefore conclude that this inference is valid. As a whole, the design choice to continue to use correct/incorrect on the final answer does impact the validity of the PSI part of TIMSS. This is addressed in the comparison inferences, where claims are made about deeper measurement of the construct of interest.

Table 6.2: Evidence for Validity Argument of PSI.

Claim	Evidence	Study	Conclusion
	Scoring inference (Figure 6.3)		
Warrant	Documents to support the scoring method	Available: Described in documentation on website	Accepted
Backing	Assessment framework that includes answers and scoring manual	Available: Described in documentation on website	Accepted
Rebuttal	Administration logs	Out of scope	Out of scope
	Generalisation inference 1 (Figure 6.4)		
Warrant	Description of construct	Study 2	Indecisive
Backing	Test matrix and test version design	Study 2	Indecisive
Rebuttal	Non response analysis	Available: Described in Study 3	Indecisive
	Generalisation inference 2 (Figure 6.5)		
Warrant	Number of items per content domain	Study 2	Accepted
Backing	Reliability analysis	Study 3	Accepted
	Extrapolation inference 1 (Figure 6.6)		
Warrant	Description of Problem Solving operationalisation	Study 2	Indecisive
Backing	Description of Task design and item analysis	Study 2	Indecisive
Rebuttal	Analysis of potential sources of variance	Study 2, Study 5, Study 6	Indecisive
	Extrapolation inference 2 (Figure 6.7)		
Warrant	Description of the construct of interest (Problem Solving)	Study 1	Indecisive
Backing	Description of procedure that involves expert judgements	Study 1	Indecisive

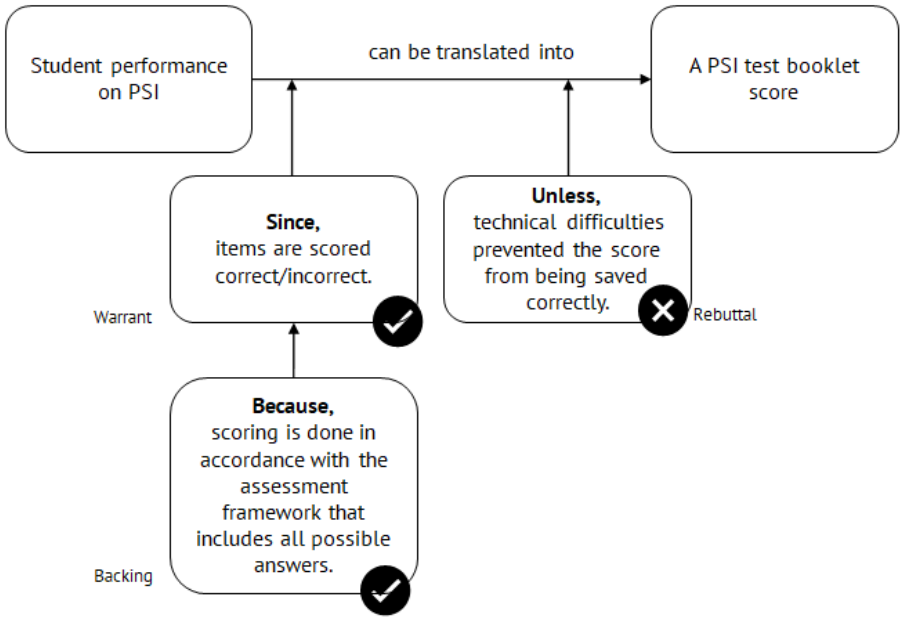

Figure 6.3: Validity Argument: PSI scoring.

6.2.2 Generalisation inference PSI

The evidence gathered for the generalisation inference for PSI is described in Study 2 (Qualitative Item Analysis) and Study 3 (Psychometric Analysis). The items included in the PSI tasks were studied and assigned to categories. These categories should subsequently match with the construct description, to decide whether the sample of items is representative for the construct. The small number of items in PSI caused some of the clusters to be underrepresented in the test matrix (Data and probability and Number). However, it is unclear to what extent this is required when problem solving is intended to be measured. Cotter et al. (2020) mention that the skew distribution of items over the test matrix of eTIMSS can be explained by the fact that PSI is intended to measure problem solving in more depth, and problem solving is more present in the higher order items. Based on the qualitative analysis that was presented in Study 2, we argue that PSI items might not all represent all aspects of problem solving. Additional research is required to draw definite conclusions about this claim. Therefore, it was decided for both the warrant and the backing to draw no conclusion on the first generalisation inference.

The rebuttal in the first generalisation inference relates to non-response. The large number of students who did not answer all items, especially in comparison to eTIMSS, is a cause for concern regarding this inference. All items are necessary in order to make claims about the construct, especially because some subdomains of the construct are only represented by a few items. Non-response would cause an even more skewed distribution of items over the content domain. However, as the studies we conducted do not provide information on whether that is the case here, it was decided not to make a definite decision on this rebuttal either. Thus, for the first generalisation inference no definite conclusions can be drawn. The evidence provided in Study 2 and 3 allow for questions regarding the validity of PSI tasks (Figure 6.4).

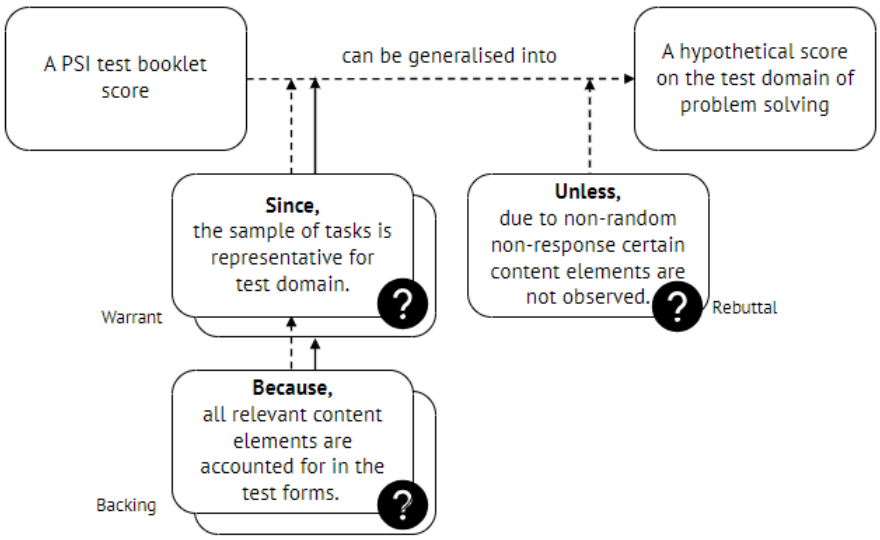

Figure 6.4: Validity Argument: PSI generalisation 1/2.

The second generalisation inference, represented in Figure 6.5, is accepted. As described in Study 2, the reliability of the PSI forms is sufficient. This led us to conclude that there

are enough items to assess students and to control for sampling error from a statistical point of view.

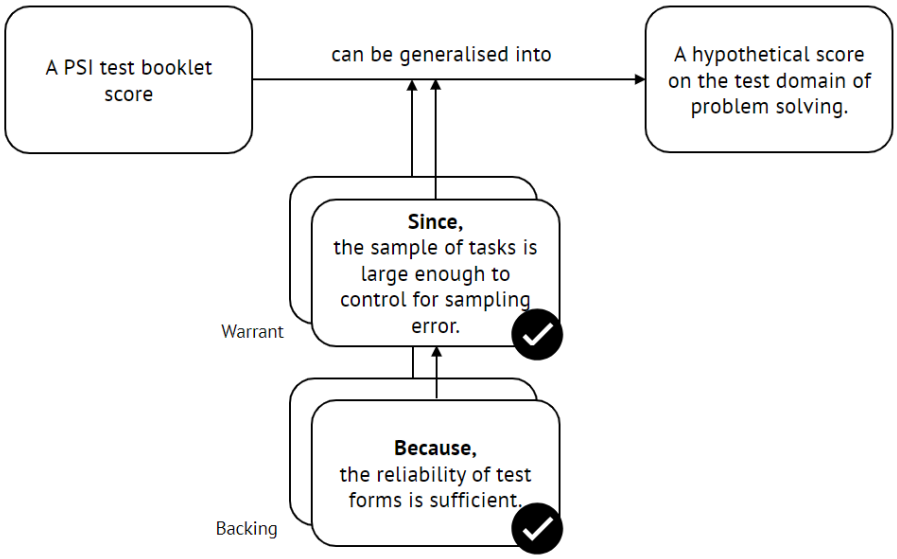

Figure 6.5: Validity Argument: PSI generalisation 2/2.

6.2.3 Extrapolation inference PSI

To evaluate the claims in the extrapolation inference (Figures 6.6 and 6.7), we relied on qualitative methods to evaluate the construct (i.e., problem solving) and the items in the PSI tasks. In Study 1 (Defining Problem Solving), the definition of problem solving as used in TIMSS was compared to other definitions of problem solving. This led us to conclude that the TIMSS definition seems to differ from other international perspectives, with the most apparent difference being that the PSI tasks are defined by an overarching theme, according to the TIMSS definition makes it problem solving. The aspects of non-routine tasks and the use of multi-step and multi-faceted solving processes are less prominent on the item level. These findings led us to conclude that the operationalisation of the construct, the actual tasks that are chosen and the description of the competence of interest could potentially be better aligned with international practices. Additional research is required to draw definite conclusions about this inference, ideally including a more extensive literature review as well as and the perspectives of more experts.

A rebuttal was included in the first extrapolation inference (Figure 6.6), which is a claim about construct irrelevant variance. In Study 2 and 3 it was found that technology-enhanced PSI items are similar to non-technology enhanced items in terms of difficulty and response times. An aspect of PSI items that could be considered distracting, and therefore leading to construct irrelevant variance, is the concept of “the umbrella” within the PSI tasks. When the overarching theme is consistent and all items within the theme are relevant and lead towards the same main goal, this aspect would in fact be supporting the problem solving character of the tasks. However, in the current PSI tasks some elements do not seem to contribute to the main goal, whereas others are somewhat confusing in terms of consistency. Additional research is required to conclude whether these aspects cause construct irrelevant variance. A definite conclusion regarding the rebuttal could therefore

not be made.

To summarise, for these inferences we conclude that the backings, warrants and rebuttal could not be accepted or rejected. Additional research is required to draw a definite conclusion about the impact on the validity of PSI.

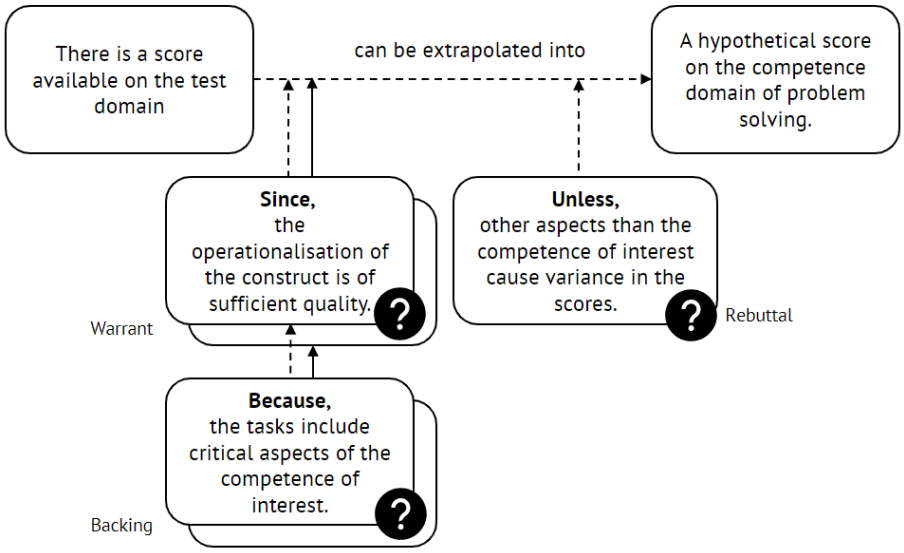

Figure 6.6: Validity Argument: PSI extrapolation 1/2.

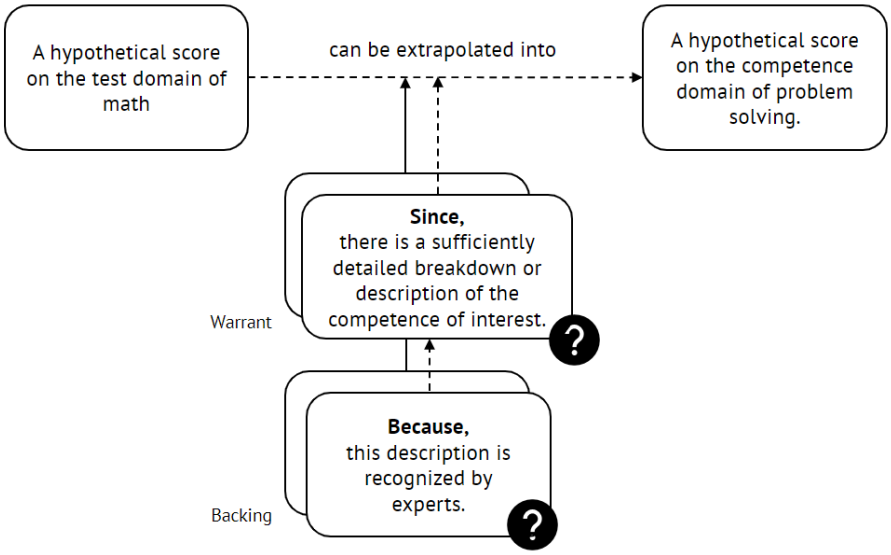

Figure 6.7: Validity Argument: PSI extrapolation 2/2.

6.3 Inferences for comparison

The inferences for comparison were included to evaluate the use of several administration modes. One overarching assumption is that conclusions about the student's proficiency can be drawn irrespective of the administration mode. However, as PSI was still in a pilot phase at the time this report was written (Martin et al., 2020), these items were not used

for the decision inference and claims are less strong regarding comparability with other administration modes. In this project, the comparability of PSI was evaluated, since it will help to understand what the potential is of these items. An overview of the inferences of comparison is provided in Table 6.3.

Table 6.3: Evidence for Validity Argument (Comparison).

Claim	Evidence	Study	Conclusion
Comparison Test Domain paperTIMSS - eTIMSS (Figure 6.8)			
Warrant	Content analysis	Study 2	Accepted
Backing	Content analysis	Study 2	Accepted
Comparison Test Domain eTIMSS – PSI (Figure 6.9)			
Warrant	Content analysis	Study 2	Rejected
Backing	Content analysis Psychometric analysis (difficulty)	Study 2 Study 4	Rejected
Comparison Competence Domain eTIMSS – PSI (Figure 6.10)			
Warrant	Content analysis Analysis of coherence of construct	Study 2 Study 4	Indecisive
Backing	Content analysis	Study 2 No process data	Indecisive

6.3.1 Comparison inference test domain paper TIMSS-eTIMSS

To draw conclusions about the first comparison inference we evaluated the comparability of paperTIMSS and eTIMSS. This comparability is taken into account in all parts of the assessment construction phase. In the digitisation process, items were not altered tremendously to ensure comparability, so not many technology-enhanced items were included. The disadvantage of this choice is that not all the possibilities of digital items were capitalised on, in the sense that the medium allowed for other measures of the construct of interest that were not included. In terms of comparability this is understandable and therefore, in this inference (Figure 6.8), both the backing and warrant are accepted.

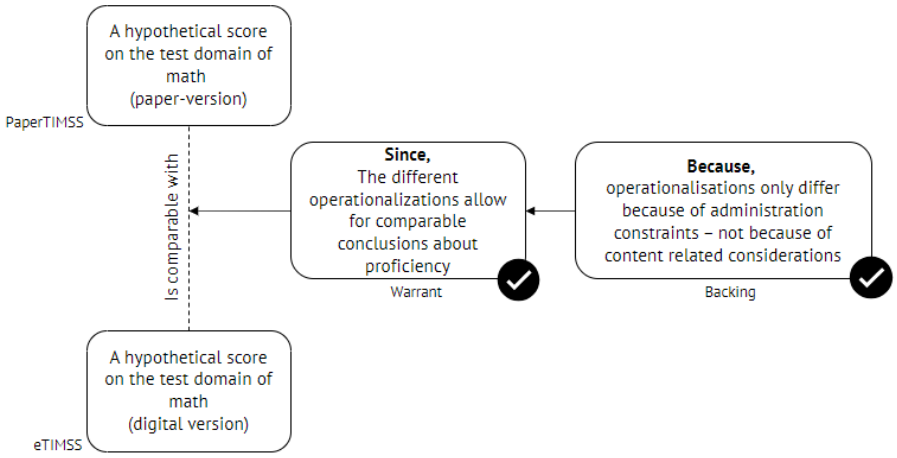

Figure 6.8: Validity Argument: Comparison Paper and eTIMSS (Test Domain).

6.3.2 Comparison inference test domain eTIMSS-PSI

When comparing the test domains of eTIMSS (math) and PSI (problem solving), it became clear that different design choices were made and that these did not lead to comparable conclusions about competence. This is by design, though: in this pilot phase different design choices were made to enable a deeper measure of problem solving. However, one could argue that the current design choice did not fully embrace the technological possibilities of technology-enhanced items. This means that the PSI tasks still include many item types that could also work on paper or in eTIMSS. Study 5 (Classical differential item functioning) showed that there were no systematic differences regarding the difficulty of items based on their technological complexity. For PSI, the technology-enhanced items are among the more easy items compared to other items. We therefore conclude that adding technology-enhanced items in itself would not impact the quality of the assessment of the competence of interest.

When drawing conclusions about the claims in this inference, we weighed in the fact that items in PSI were chained together to form a larger problem and that some items in the PSI tasks included non-routine solutions. Therefore, we conclude that the operationalisation for both eTIMSS and PSI seem to differ on more aspects than just administration related aspects. Therefore, both the warrant and backing are rejected, as is visualised in Figure 6.9.

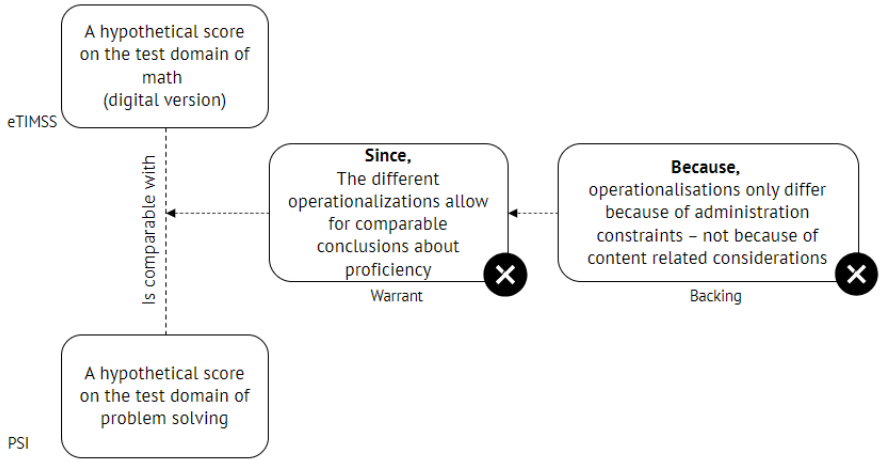

Figure 6.9: Validity Argument: Comparison eTIMSS and PSI (Test Domain).

6.3.3 Comparison inference Competence Domain eTIMSS-PSI

The final comparison concerns the competence domain that is assessed in eTIMSS and PSI. Documentation regarding PSI states that these tasks intend to measure problem solving on a deeper level, but that it is still the same construct. This was not described in sufficient detail to be able to accept these claims purely based on this statement. Study 4 (Comparison eTIMSS and PSI) was conducted to obtain more evidence for this claim. The findings of this study suggest that different skills are required to solve eTIMSS items than to solve PSI items. Also, it was shown that for PSI the items seem more suited to above-average performing students, which is not the case for eTIMSS items. This finding also indicates that there are differences in the construct and operationalisation of PSI and eTIMSS. When

taking this evidence into account, we come to the conclusion that there is evidence that PSI and eTIMSS do not measure exactly the same construct, but it remains unclear whether PSI is a deeper measure of problem solving. One design choice that is not supportive of this claim is the fact that PSI tasks are scored as correct or incorrect, whereas a deeper measure of problem solving could potentially be obtained by observing strategies that students use to solve items. This would, however, also require items that aim to provoke the use of different strategies.

All in all, the evidence that we found regarding the competence domain of math and whether a deeper measure of problem solving was assessed in PSI was not convincing. Therefore, it was decided that additional research is required to be able to draw a definite conclusion about this inference. This is visualised in Figure 6.10.

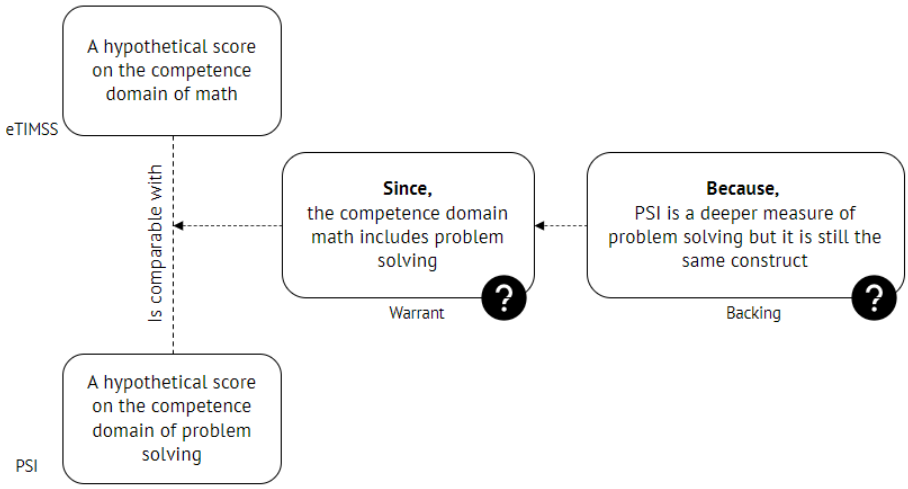

Figure 6.10: Validity Argument: Comparison eTIMSS and PSI (Competence Domain).

7. Conclusion and discussion

In this project we used an argument-based approach to validation to investigate the impact of the use of technology-enhanced items in international large-scale assessments. Technology-enhanced items are thought to yield better measures of the construct of interest and to allow drawing in-depth conclusions on proficiency. At the same time, different aspects regarding the use of technology-enhanced items are not clear.

In this particular project we chose to focus on the use of technology-enhanced items in the TIMSS 2019 mathematics cycle, in which about half of the countries opted for a digital administration of the assessments (eTIMSS). In addition, we chose to restrict this study to the construct of mathematics and took a deep dive into the newly developed PSI (problem solving and inquiry) tasks.

This chapter draws conclusions about the use of technology-enhanced assessment and their relation to the validity of the results of TIMSS 2019. We provide suggestions for the design and use of these items for a next cycle. Finally, we reflect on this study and especially it's limitations.

7.1 Conclusions from quantitative studies

From the psychometric analyses we can conclude that the PSI items demonstrated psychometric properties (reliability, factor structure, and amount of DIF) that are comparable to the more established eTIMSS items. This is promising as it indicates that, at least from a psychometric perspective, interactive assessment enables standardized and controlled educational measurement with comparable psychometric quality as the more traditional assessments. There are, however, some challenges. First, as the psychometric quality of technology-enhanced items is comparable to the more traditional way of assessment, their added value consequently relies on their improved validity and/or improved respondent engagement. With respect to the latter, however, the relatively high number of non-responses

and the relatively low accuracy of the PSI items both indicate that respondents do engage with these items differently than (at least) the eTIMSS items, but not necessarily in the desired way. A second challenge arises in the process data of these items. As discussed, process data appear promising to reveal differences in response strategies. On the one hand, we found successful applications of the on-screen times in the present project to identify intra-individual differences and explanations for traditional DIF effects. However, the effects found in the process data are relatively heterogeneous, making them difficult to interpret in general and requiring an interpretation that is more specific to the country (and even respondent). In the present project it cannot be ruled out that the on-screen process data mostly reflect effects of motivation; that is, in different analyses we found that respondents sped up at the cost accuracy, which may suggest a lack of motivation or giving up easily. For the process data concerning the use of the calculator it can also not be ruled out that the main effect (use of the calculator strongly correlated with performance) is related to motivation. This is an interesting result in itself, as it demonstrates the validity of the TIMSS process data as a measure of test engagement (e.g. Wise and Kong, 2005). However, the we may be able to extract more information in the future. Similarly to the screen times (on which we focused mainly), most of the process data in TIMSS is collected per screen instead of per item. To improve the resolution of the inferences from the eTIMSS and PSI process data, data should ideally be collected in an item-wise controlled way (e.g. one item per screen, no screen revisits). Of course, this may be undesirable or unpractical for different reasons (e.g. clustered items, time management of the respondent), but it may be worth considering if the main interest is in studying response processes.

Differential item functioning is unavoidable in large-scale assessments. It is thus not surprising that we did find DIF in four booklets from the eTIMSS and the items from the PSI scale. However, remarkably, the PSI scale did not show a substantially different pattern of DIF as compared to the eTIMSS, indicating that the PSI items are performing psychometrically equally well as the eTIMSS in terms of group comparisons. A difference that should be noted is that for the eTIMSS, the technology-enhanced items are comparable to the more traditional items, while for the PSI items, the technology-enhanced items are the easier items.

In the DIF analysis of the response times of the technology-enhanced items and the non-enhanced items, no systematic differences were found for eTIMSS. For PSI, we found that DIF for TE screens across countries was not systemically different from non-TE screens. The two items with the longest response times were technology-enhanced, but these both required multiple mathematical actions. Therefore, for both eTIMSS and PSI it can be concluded that technology-enhanced items do not measure computer skills instead of mathematical skills.

In the analysis of intra-individual differences, the response times were shown to be useful in flagging two items for which respondents slowed down to obtain a higher score. These differences seem related to the difficulty level of the items and the item setup, which may be related to motivation issues: Less motivated respondents responded quickly and relatively inaccurately so that the slower responses are relatively more accurate. The heterogeneity of the effects across countries was large, making the interpretation of the effects highly country specific.

Using the time variables it was also possible to explain some of the DIF found in Study

5. That is, for various items, the DIF was found to be related to item specific differences in timing. Again, due to the heterogeneity of the effects across countries, the interpretation of the effects depends on the specific country, as was evident from the results.

Finally, the use of a calculator was found to be a strong predictor of accuracy and time spent on the item. As calculator use differs strongly across countries, there are possibly cross-country differences in educational practices with respect to calculator use in class. Alternatively, it is also possible that the use of the calculator is an indicator of motivation, as respondents that use the calculator performed well on the items, and respondents that did not use the calculator scored much lower.

7.2 Conclusions from qualitative studies

The conclusions that can be drawn from the quantitative analyses show that the use of technology-enhanced items does not lead to big differences in terms of psychometric properties or DIF. However, based on the qualitative analyses we conclude that the opportunities offered by technologically-enhanced items are not fully deployed yet. We recognize that technological restrictions played a part in these design choices. Items in the PSI tasks were often multiple-choice or open-ended questions, even though there are many more technological possibilities. Due to the limited technological possibilities that were employed, it remains unclear whether more technologically advanced items would make a good fit or even a better measure of the construct of interest for international large-scale assessments.

One of the reasons to introduce technology-enhanced items was their potential to assess complex skills more accurately. PSI was intended to be a better measure of mathematical or scientific problem solving. Based on the qualitative analyses, we argue that it is questionable whether that goal was met. This is partly because of the definition and operationalization that was chosen, not due to the items. Using the 'umbrella' approach where multiple items are brought together under one theme is, in itself, not an operationalization of problem solving. Furthermore, this design choice leads to 'rider' items where items are dependent on each other. In some instances, an effort is made to make items independent of previous items, but due to the joint theme of these items, this could be confusing for students.

Finally, the items included in the current cycle do not encourage candidates to engage with the items in such a way that the process data can be used to draw inferences about the strategies they used. Some items require multiple (digital) steps, but these steps are often not mandatory to solve the item. It would be interesting to experiment with items that encourage or require more varied strategy use. And also, to design studies where the design principles of these items can be tested to investigate whether these items are actually able to provoke the intended behavior.

7.3 Technology-enhanced items and validity

We conclude that the addition of elementary technology-enhanced items does not threaten the validity of the results of international large-scale assessments. However, it also does not add to the validity of the results when more complex skills are to be assessed. With the current items and the design choices made in the PSI tasks it is not possible to establish whether technology-enhanced items could be used in the context of international large-scale assessments to measure problem solving on a deeper level.

To be able to study the impact of these innovative items in the future, some recommendations regarding their design can be made. First of all, these items should be designed with the aim to provoke strategies. This means that these items would require a strategy to be solved, but also, that it is possible to use different strategies. Subsequently, these different strategies should lead to different behavior in the assessment environment so that they can be observed in the data.

Secondly, we recommend that items use more technological possibilities to obtain the answers of students (drawing, typing, drag and drop, rotating figures, etc). Alongside these technological possibilities, items that involve a multi-step solution process and items with multiple degrees of freedom in finding the solution can be included. It could even be possible to include items for which several answers are correct.

And finally, we recommend that the items themselves contain elements of problem solving by including non-routine tasks. These could be used instead of the umbrella with the overarching problem that is solved by walking through pre-defined steps and answering items along the way.

7.4 Limitations

This study aimed to use and develop advanced data analytics to establish the validity of the results of international large-scale assessments that include technologically-enhanced items. However, due to the lack of complex technology-enhanced items and the specific design choices of the PSI tasks this did not deliver insightful results.

The analyses that were performed on the response times show differences in groups. However, and this is a more general limitation of international large-scale assessments, these differences seem to be caused by differences in motivation where, ideally, one would like to draw conclusions about strategy use or proficiency.

During this study choices were made regarding the scope of this project. We focused on the math part of TIMSS 2019, and only one PSI task was studied in depth in our qualitative studies. Furthermore, the qualitative analysis of both Study 1 and 2 could be set up on a larger scale, which would allow for definite conclusions regarding the validity of PSI tasks. In a high-level analysis of the other content of TIMSS 2019, it was established that the items that were included in this study were representative of the other content in terms of technology use. Therefore, conclusions drawn about the technology-enhanced items can be generalized to all items.

Finally, we took an argument-based approach to validation. This approach includes an assessment of risks and concerns which is guided by choices made by the researchers. Another team could have placed different accents. Also, when much information is available, it might always be possible that counter-evidence exists that is not taken into account. Therefore, we stress that validity efforts are never finished and that understanding of the validity of assessment results should be an ongoing effort in international studies.

Bibliography

Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance.

Cotter, K. E., Centurino, V. A., & Mullis, I. V. (2020). Developing the TIMSS 2019 mathematics and science achievement instruments. *Methods and procedures: TIMSS*, 1–1.

Feskens, R., Fox, J.-P., & Zwitser, R. (2019). Differential item functioning in PISA due to mode effects. In *Theoretical and practical advances in computer-based educational measurement* (pp. 231–247). Springer, Cham.

Fishbein, B., Martin, M. O., Mullis, I. V., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education*, *6*(1), 1–23.

Hamhuis, E. R., Glas, C. A., & Meelissen, M. R. (2018). De etimss equivalentiestudie: Van papieren toets naar tablettoets, zijn er verschillen?

Hastedt, D., & Rocher, T. (2020). International large-scale assessments in education: A brief guide. IEA compass: Briefs in education. number 10. *International Association for the Evaluation of Educational Achievement*.

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: How big is the 'mode effect' and what has been done about it? *Oxford Review of Education*, *44*(4), 476–493.

Johansone, I. (2020). Survey operations procedures for TIMSS 2019. In *Methods and procedures: TIMSS 2019 technical report.* (pp. 6.1–6.28). International Association for the Evaluation of Educational Achievement, Amsterdam. https://timssandpirls.bc.edu/timss2019/methods/chapter-6.html

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448–457.

Kane, M., & Brennan, R. (2006). Educational measurement. *Validation (4th ed., pp. 17–64)*. *Westport: American Council on Education/Praeger*.

Lindquist, M., Philpot, R., Mullis, I. V., & Cotter, K. E. (2017). TIMSS 2019 mathematical framework. In *TIMSS 2019 assessment frameworks.* International Association for the Evaluation of Educational Achievement, Amsterdam.

Lockheed, M. E., & Wagemaker, H. (2013). International large-scale assessments: Thermometers, whips or useful policy tools? *Research in Comparative and International Education*, *8*(3), 296–306.

Martin, M. O., Mullis, I. V., & Foy, P. (2017). TIMSS 2019 assessment design. In *TIMSS 2019 assessment frameworks.* International Association for the Evaluation of Educational Achievement, Amsterdam.

Martin, M. O., von Davier, M., & Mullis, I. V. (2020). Methods and procedures: TIMSS 2019 technical report. *International Association for the Evaluation of Educational Achievement*.

MeasuredProgress & ETS. (2012). Technology enhanced item guidelines.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research*, *13*(2), 127–143.

Mullis, I. V. (2017). Introduction. In *TIMSS 2019 assessment frameworks.* International Association for the Evaluation of Educational Achievement, Amsterdam.

Mullis, I. V., Martin, M. O., Fishbein, B., Foy, P., & Moncaleano, S. (2021). Findings from the TIMSS 2019 problem solving and inquiry tasks.

Mullis, I. V., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). TIMSS 2019 international results in mathematics and science. *Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-results*.

Robitzsch, A., & Lüdtke, O. (2020). A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psychological Test and Assessment Modeling*, *62*(2), 233–279.

Schwippert, K., & Lenkeit, J. (2012). *Progress in reading literacy in national and international context. the impact of PIRLS 2006 in 12 countries*. Waxmann Verlag.

Sireci, S. G., & Zenisky, A. L. (2011). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In *Handbook of test development* (pp. 343–362). Routledge.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models.

Toulmin, S. (1953). *The philosophy of science*. Hutchinson London.

Toulmin, S. (2003). *The uses of argument*. Cambridge university press.

Verhelst, N. D. (2012). Profile analysis: A closer look at the PISA 2000 reading data. *Scandinavian Journal of Educational Research*, *56*(3), 315–332.

von Davier, M., Foy, P., Martin, M. O., & Mullis, I. V. (n.d.). Examining eTIMSS country differences between eTIMSS data and bridge data.

Wagemaker, H. (2013). International large scale assessment (ILSA) programs and the challenges of consequential validity. *Validity and test use: An international dialogue on educational assessment, accountability and equity*, 217–233.

Wagemaker, H. (2020). *Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement*. Springer Nature.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

Wools, S., Eggen, T. J., & Béguin, A. A. (2016). Constructing validity arguments for test combinations. *Studies in educational evaluation*, *48*, 10–18.

Wools, S., Molenaar, M., & Hopster-den Otter, D. (2019). The validity of technology enhanced assessments—threats and opportunities. In *Theoretical and practical advances in computer-based educational measurement* (pp. 3–19). Springer, Cham.

Wools, S., Sanders, P., & Eggen, T. (2010). Evaluation of validity and validation by means of the argument-based approach. *Evaluation of Validity and Validation by Means of the Argument-based Approach*, 63–82.

Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika*, *82*(1), 210–232.

List of Figures

List of Tables