Comparing apples with oranges? An approach to link TIMSS and the National Educational Panel Study in Germany via equipercentile and IRT methods
Nissen, Annika; Ehmke, Timo; Köller, Olaf; Duchhardt, Christoph

Published in:
Studies in Educational Evaluation

DOI:
10.1016/j.stueduc.2015.07.003

Publication date:
2015

Document Version
Peer reviewed version

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Comparing apples with oranges? An approach to link TIMSS and the National Educational Panel Study in Germany via equipercentile and IRT methods

Annika Nissen¹,*, Timo Ehmke ², Olaf Köller³, Christoph Duchhardt⁴

¹ Leuphana University of Lüneburg, Schumannstraße 1, 21335 Lüneburg, Germany
² IFN – Leibniz Institute for Science and Mathematics Education, Oshausenstraße 62, 24118 Kiel, Germany
³ Centre for International Student Assessment, Affiliated Institute of the Technical University Munich, Marsstraße 20-22, 80333 Munich, Germany
⁴ University of Bremen, Bibliothekstr. 1, 28359 Bremen

ABSTRACT

The aim of this linking study is to classify the National Educational Panel Study (NEPS) according to the criterion based reference frame of the Trends in International Mathematics and Science Study (TIMSS). One requirement for a strong linking is a high conceptual overlap of the frameworks. The results of the comparison between NEPS and TIMSS indicate a high overlap. To link the studies the equipercentile equating and an IRT linking approach are compared. The results show that both methods showed similar descriptive statistics and satisfying classification consistency. Nevertheless, the equipercentile equating has small advantages and is therefore more suitable. Thus, inferences are possible on a population level, but the results should not be reported or interpreted on an individual level.

© 2015 Published by Elsevier Ltd.

1. Introduction

There are many national and international large-scale assessment studies which measure mathematics competencies of primary and secondary school students as, for example, the Trends in International Mathematics and Science Study (TIMSS; Mullis, Martin, Ruddock, O’Sullivan & Preuschoff, 2009) and the National Educational Panel Study (NEPS; Blossfeld, von Maurice & Schneider, 2011) in Germany. Increasingly, national and international assessments have been linked with the aim to explain differences in the results of the corresponding studies (e.g., Gronno & Olsen, 2007; Neidorf, Binkley, Gattis & Nohara, 2006; Nohara, 2001; Wu, 2010) or to use the benchmarks in other national studies (e.g., Cartwright, Lalancette, Mussio & Xing, 2003; National Center for Education Statistics, 2013; Phillips, 2007).

In Germany, the Standing Conference of the Ministers of Education and Cultural Affairs of the Laender in the Federal Republic of Germany (KMK) calls for a link between national and international assessment studies (KMK, 2006). However, it is not feasible to simply compare the assessments and to relate the results accordingly for several reasons. Amongst others, the assessments are based on different frameworks and the results are not reported on the same scales. To enable a valid comparison of the findings, instruments from both studies have to be linked to a common scale.

This study examines a linkage between the mathematics tests from TIMSS 2011 and NEPS 2010 that were developed for the same age cohort (students at the end of grade 4/beginning of grade 5). NEPS is a longitudinal study in Germany which analyzes educational processes over the lifespan. In NEPS, no proficiency levels are defined. So the results of NEPS cannot be interpreted in a criterion-based manner. TIMSS 2011 is an international study which assesses mathematics and science achievement of students with the overarching goal of improving teaching and learning. In TIMSS, achievement is reported at four proficiency levels (international benchmarks). These benchmarks describe what students typically know and can do in mathematics (Mullis, Martin, Foy & Arora, 2012). Linking NEPS with TIMSS has the advantage that the results of NEPS could be classified in the international and criterion-based reference frame of TIMSS. By this

* Corresponding author.

E-mail addresses: annisen@leuphana.de (A. Nissen), tehmkke@leuphana.de (T. Ehmke), koeller@uni-kiel.de (O. Köller), christoph.duchhardt@uni-bremen.de (C. Duchhardt).

http://dx.doi.org/10.1016/j.stueduc.2015.07.003
0191-491X/© 2015 Published by Elsevier Ltd.
means, linking could extend the interpretations of the NEPS scores. For example, students who do not exceed the TIMSS low international benchmark, or who reach the advanced international benchmark, could be examined longitudinally within NEPS.

Accordingly, 733 fourth graders took both test forms, the TIMSS mathematics test for the end of grade 4 and the NEPS mathematics test for the beginning of grade 5. Due to the fact that “different methods and different groups do not produce identical ‘equatings’” (Lord & Wingersky, 1984, p. 455), we compared two different linking approaches. Furthermore, different linking methods are based on different assumptions. Hence, using two methods and comparing the results is a form of quality control (Kolen & Brennan, 2010). In this study, the equipercentile equating and item response theory (IRT) scale transformation were applied to find out which method fits the data better concerning (1) descriptive statistics such as means, standard deviations, skewness, and kurtosis and (2) classification accuracy according to the TIMSS international benchmarks. One requirement for a strong linking is a high conceptual overlap of the studies, while the utility and reasonableness of a linking can be influenced by the degrees of similarity (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Kolen & Brennan, 2010; Linn, 1993; Mislevy, 1992). Therefore, the frameworks and test specifications of the two studies will be compared beforehand.

In Section 2, after a short description of the TIMSS and NEPS frameworks and test specifications, an introduction to the linking methods will be given. Afterwards, we will focus on the current state of research, giving an overview of some studies which concentrate on different aspects of linking. In Section 3, the research questions will be formulated. In Section 4, the methods of the linking-study will be presented. Section 5 is devoted to the results. In a first step, we will look at common features and differences in the frameworks and test specification of the two studies. After this, the results of the linkings will be presented, first for the equipercentile linking and then for the IRT linking. Afterwards, we will compare the outcomes, going on to demonstrate the results of the conversion. In Section 6, the results will be interpreted and discussed.

2. Theoretical background

2.1. About TIMSS

In 1995, the Third International Mathematics and Science Study (TIMSS; current: Trends in International Mathematics and Science Study; Mullis et al., 2009) was conducted by the International Association for the Evaluation of Educational Achievement (IEA) for the first time. TIMSS 2011 is a cross-sectional study which assesses mathematics and science achievement of students in grades 4 and 8. As it was repeatedly conducted after 4 years, it created the opportunity to measure trends across age cohorts in the educational systems. The overall aim of the study is to compare the countries’ outputs of their educational system and to identify the influencing factors. Thus, it means to explain the various countries’ differing results, and to identify possibilities which could support the improvement of the respective educational systems. In 2011, more than 60 countries participated in TIMSS.

In addition to the assessment instruments in mathematics and science, students’ social background and other characteristics are collected via questionnaires. The basis for the frameworks of the tests is formed by the curriculum model which consists of three aspects: (1) what the students in the countries are expected to learn (intended curriculum); (2) what the students actually learn in the classrooms (implemented curriculum); and (3) the final outcome (attained curriculum).

2.1.1. TIMSS 2011 – mathematics test for fourth graders

The TIMSS 2011 assessment for fourth graders took 72 min and contained 177 mathematics items and 175 science items. The participation for the selected sample was obligatory. The items were included in 14 booklets, each one with four blocks, i.e., two mathematics and two science blocks. IRT was used to scale the data and develop the reporting scales. For the constructed responses (46.9%) a two-parameter model was used, for the multiple-choice items (53.1%) a three-parameter model was applied. Additionally, a partial credit model was employed for nine items. In order to estimate the students’ personal proficiency, the test instruments of TIMSS apply the plausible value approach (Foy, Brossman & Galia, 2012).

The mathematical framework differentiates two domains, the content and the cognitive domain. The content domain includes the subdomains number (50%), geometric shapes and measures (35%), as well as data display (15%). The cognitive domain differentiates between knowing (40%), applying (41%), and reasoning (19%) (Foy et al., 2012).

2.2. About NEPS

The National Educational Panel Study (NEPS) is a longitudinal study in Germany. The aim is to measure competencies over the lifespan (from early childhood to late adulthood) and describe the education processes and the developmental trajectories (Blossfeld et al., 2011). The main interest is to measure the development of competencies and ascertain possible aspects of impact. Therefore, NEPS uses a multi-cohort sequence design. There are six cohorts, each starting at different stages of transition in the educational system: newborns, four-year-olds, fifth graders, ninth graders, university freshmen to higher education students, and adults aged between 25 and 65 years. Four starting cohorts began in 2010, one in 2009, and one in 2012. The participants will be followed across their lifespan.

NEPS was initiated and funded by the German Federal Ministry of Education and Research (Blossfeld, von Maurice & Schneider, 2009). For the long-term organization of the study, the Leibniz Institute for Educational Trajectories (LIfBi) was established in 2014. NEPS assesses four competence areas (Weinert et al., 2011): (1) domain-general cognitive abilities and capacities; (2) domain-specific cognitive competencies (German language, mathematics, and science competencies); (3) meta-competencies and social competencies; and (4) stage-specific attainments, skills, and outcome measures.

2.2.1. NEPS 2010, grade 5 – mathematics test

In our equivalence-study, we used the NEPS mathematics test for fifth graders implemented at the beginning of their school year. The participation was optional. The NEPS mathematics test is based on the concept of mathematical literacy, known, for example, from PISA (OECD, 2010). The framework differentiates between two domains: the content areas as well as mathematical and cognitive processes. The content areas include the subdomains quantity (33%), change and relationship (25%), space and shape (21%), and data and chance (21%). The mathematical and cognitive processes differentiate between argumentation, communication, modeling, problem solving, representing, and applying technical skills (Weinert et al., 2011).

For fifth graders, the test takes about 30 min including 25 items. The NEPS test is scaled based on the IRT (Duchhardt & Gerdes, 2012). A one-parameter Rasch model is used. In NEPS, students’ Weighted Likelihood Estimates (WLE) scores are calculated as estimates for the students’ achievement scores. The mathematics items are multiple choice (13 items), complex multiple choice
(1 item), or constructed response (11 items). 24 items are dichotomous and one is partial credit.

2.3. Linking-methods

Linking became relevant through the increased significance of using different test forms to guarantee test security (Kolen & Brennan, 2010). Using the same test for two groups at different points in time could cause a violation of test security because exchange between the groups could take place. Testing one group at two points in time could cause memory effects and influence test security, too. To avoid this, alternate (also called parallel or equivalent) test forms were constructed. These test forms should be equivalent concerning the content, the statistical specifications, and the administrative conditions (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999). However, it is also possible that the test forms vary in difficulty. To correct these differences, technical methods of equating can be used. Because of the strict guidelines of equating, the need for weaker forms of linking with less rules and more possibilities arose (Pommerich, 2007). In the following, these different forms of linking are described.

Mislevy (1992) and Linn (1993) differentiate between four types of linking: equating, calibration, projection, and (statistical or social) moderation. Equating is the strongest type of linking. It is often used if parallel tests exist. Therefore, the test blueprint and the population must be equal. The equipercentile and linear equating method can be used (see Section 2.3). If the tests differ in purpose and test characteristics, then the scores can be linked by calibration (e.g., vertical scaling). To link two tests IRT methods can be applied, given that the sample size is large enough. Projection is the third type of linking according to Mislevy (1992) and Linn (1993). It is unidirectional, this means the results of one test X can predict the results of another test Y; but if Y is used to predict the results of X different outcomes are expected. Regression analysis is conventionally used as a method to predict the results. The weakest form of linking is moderation. This can be used only for very general comparisons (Feuer et al., 1999). Here, the test specifications and the population can differ.

The linking-variations could be located on a linking continuum. In the approaches of Mislevy (1992) and Linn (1993) there are weaker and stronger kinds of linking. The stronger the linking, the stronger the conclusions that can be drawn. Concerning this, the extent to which the tests agree on, e.g., test specifications and frameworks, is also important.

Kolen and Brennan (2010) describe another approach to linking test forms. They specify terms of degrees of similarity, whereby equivalence analysis should include four features: (1) inferences (e.g.: Which kind of inferences could be drawn from the studies?); (2) construct (e.g.: Do the two tests measure the same construct?); (3) populations (e.g.: Are the tests created for the same target group?); and (4) measurement characteristics (e.g.: Are the conditions for the test implementation similar?).

2.4. Current state of research

In equivalence studies, the similarities and differences between two or more tests have to be examined before the tests are linked. The degree of similarity influences the linking type and the linking method which can be applied. Depending on the goal of the studies, different aspects are analyzed. Usually, there are three different research purposes:

2.4.1. (1) Comparing studies to explain differences in outcome

One research purpose is to show differences and commonalities between assessments in order to gather more information with the aim of improving the interpretation of the studies’ results and of explaining the differences in outcomes. For instance, Neidorf et al. (2006) compared the frameworks and the items of the National Assessment of Educational Progress (NAEP), TIMSS, and PISA 2003. They found that NAEP has more similarities with TIMSS than with PISA, and recorded the main differences of the studies. They stated that more in depth analysis of the items would reveal more important differences. Overall, they concluded that the three assessments are complementary. Hence, one study might provide more information for one specific topic or skill than the other studies. Wu (2010) compared the survey methodologies and frameworks of PISA 2003 and TIMSS 2003. She used the linking method projection and analyzed the data by means of regression analysis. One of her conclusions was that content balance and sampling definitions affect the outcomes at country level. Furthermore, Wu found out that the PISA assessment instruments use more words than the TIMSS tests and assumes that students with lower reading competencies will perform unequally in TIMSS and PISA.

2.4.2. (2) Comparing studies to explain differences in benchmarks/proficiency levels

Another research purpose is to compare one or more studies in relation to differences in national and international benchmarks. Thus, it can be examined if the requirements in the national study are higher or lower than by international comparison. This was, for example, the research purpose of Hambleton, Sireci and Smith (2009). They compared NAEP with TIMSS and PISA in 2003, questioning whether the proficiency levels in NAEP are set too high. They concluded that, in fact, this is not the case in an international context.

2.4.3. (3) Linking tests to locate the outcomes of the national studies in an international reference

Linking a national test with an international test provides the opportunity to locate the outcomes of the national study on the international scale. As a consequence, it is possible, e.g., to examine the relative strengths and weaknesses of a country in an international comparison. Another possibility is to use the international benchmarks and with it a criterion reference in a national study. For instance, Cartwright et al. (2003) linked British Columbia’s annual Foundation Skills Assessment (FSA) with PISA 2000 to compare and report the results on a common scale. They used the linking method concordance (statistical moderation). One of their outcomes was that a transformation of the FSA results to the PISA scale would lead to a higher percentage of students at the top reading level.

The most recent linking study of NAEP and TIMSS was conducted in 2011 (National Center for Education Statistics, 2013). The aim was to compare the competencies of U.S. students with the competencies of students in other countries, with the result that mathematics competencies of students in 36 states were higher than the TIMSS average of 500; furthermore, the students of 47 states had higher science competencies compared to the average of 500 in TIMSS. The results are based on the method of statistical moderation, though the results of the projection and calibration method are approximately the same.

3. Research questions

High conceptual overlapping of the studies is one requirement for a strong linking (Feuer et al., 1999; Kolen & Brennan, 2010; Linn, 1993; Mislevy, 1992). Thus, this study investigates the following research questions:
(1) How equivalent are the mathematics assessments in TIMSS 2011 grade level 4 and NEPS grade level 5 concerning their inferences, constructs, populations, and measurement characteristics?

(2) Are the mathematical content areas comparable between the TIMSS and the NEPS test?

(3) Are the mathematical items of the two studies comparable regarding formal and language demands?

Furthermore, the study also takes a look at two different linking methods (classical test equating and IRT).

(4) Which linking method (equipercentile or IRT) should be preferred regarding the descriptive statistics such as means, standard deviations, skewness, and kurtosis and the classification consistency according to the TIMSS international benchmarks?

4. Method

4.1. Data collection

The international TIMSS 2011 assessment and the German NEPS are two separate studies that share neither common test items nor common grade levels (end of grade 4 vs. beginning of grade 5). However, to link these two large-scale assessments, a validation study was conducted using a single group design.

The linking study was arranged as a two-day assessment at schools which participated in a program for school development. On the first day, every student completed one of the 14 TIMSS booklets (each with two mathematics and two science blocks). On the second day, the students completed the NEPS mathematics tests for 5th graders. The single group design allows a linking of the scales between the TIMSS and NEPS mathematics assessments.

The test booklets used in the linking study are the same as in the main studies of TIMSS 2011 (TIMSS2011) and NEPS starting cohort 3, grade 5 (NEPS2010). Therefore, it is possible to link the assessments scales (A) between the linking study TIMSSLink and the international TIMSS2011 study, and (B) between the linking study NEPSLink and the main NEPS2010 (see Fig. 1). Since the student populations are different but the test instruments are the same, the data collection design can be classified as Non-Equivalent Groups with Anchor Test Design (NEAT; from von Davier, Holland & Thayer, 2004). The linking study provides the opportunity to (C) link NEPSLink and TIMSSLink, because the same students took both mathematics tests (single group design).

The data set realized in the linking study consists of 1581 fourth graders who took the TIMSS test and 752 fourth graders who took the NEPS test from 78 German primary schools in altogether 80 classes. Overall N = 733 fourth graders (52% male, 48% female) took both the NEPS and the TIMSS mathematics test. The study was conducted in the spring/summer of 2011.

4.2. Conceptual overlapping of TIMSS and NEPS

First, a framework review was conducted. The two studies have been compared regarding their degrees of similarity including four features: inferences, construct, populations, and measurement characteristics (Kolen & Brennan, 2010). Additionally, two expert reviews were carried out for an in-depth analyses of the conceptual overlapping.

1. Expert review: Three assessment specialists rated the 24 NEPS mathematics items according to the TIMSS 2011 assessment framework. Therefore, all 24 NEPS items are classified as belonging to one of the three content domains (number, geometric shapes, and measures, as well as data display) and to one of the three cognitive domains (knowing, applying, and reasoning). To control for rater agreement, the percent agreement (pa) and the Cohen's kappa (κ) were calculated. The median of the percent agreement is pa = 72% and the Cohen's kappa is κ = .44. Thereby, the rater agreement in the first expert review is acceptable (Wirtz & Caspar, 2002).

2. Expert review: Three experts rated the 24 NEPS and the 177 TIMSS mathematics items according to formal and language demands. For this purpose, the items were classified regarding (1) the presence of a picture/photo, table, graph, formula, and the response format; and (2) the number of words, sentences, and mathematical terms. Again, to control for rater agreement the percent agreement (pa) was calculated. The median of the percent agreement is pa = 87%. Additionally, the Cohen's kappa (aspects of (1)) and the Intra-Class-Correlation (ICC: aspects of (2)) were calculated. The Cohens kappa is κ = .47 and the ICC = 1.0. Thus, the rater agreement in the second expert review is acceptable, too (Wirtz & Caspar, 2002).

4.3. Scoring and data procedures

The scoring of the TIMSS 2011 mathematics data in the linking study was conducted by applying the international coding rules from TIMSS 2011 (Martin & Mullis, 2012). The test scores were computed using a three-parameter IRT model with fixed item parameters from the international database of TIMSS 2011 (Martin & Mullis, 2012). WLE scores were then converted into the international TIMSS 2011 achievement scale metric by using the transformation equation provided in Martin and Mullis (2012).

The scoring of the 24 mathematics items of the grade 5 assessment was done by using the NEPS coding rules. The raw scores were computed based on a one-parameter Rasch model with fixed item parameters taken from the NEPS 2010 main study (Duchhardt & Gerdes, 2012). Then, students’ WLE scores were linearly transformed to achieve only positive integer values.

4.4. Linking procedures

The aim of this study was to link scales from TIMSS 2011 and NEPS 2010, maintaining the international TIMSS reporting scale. Many different methods and tools exist to link or equate two or more tests depending on which linking form and data collection design is used. For this reason, we will compare the results of two different methods: the equipercentile equating versus an IRT linking.

A generally applied equating method is the equipercentile equating (e.g. Cartwright, 2012; Cartwright et al., 2003; Hambleton et al., 2009). If the equipercentile method is used to link or equate two tests, the percentile ranks for the score distributions have to be
determined. After this, the scores of the two tests with the same percentile rank are declared equivalent (Kolen & Brennan, 2010; Muraki, Hobo, & Lee, 2000). For example, if score 400 is equal to percentile rank 15 on the TIMSS scale, and score –1.5 is equal to percentile rank 15 on the NEPS scale, then score 400 and score –1.5 on the NEPS scale are considered as equipercentile equivalent scores. The equipercentile equating in our study was done by using the computer software LEGS 2.01 (Brennan, 2003), which allows linking with equivalent or single group designs.1

IRT linking methods focus on modeling students’ responses to items instead of using the score distributions as a linking basis. Using IRT methods to equate two or more test forms typically requires three steps (Kolen & Brennan, 2010): (1) estimating the item parameters (e.g., using PARSCALE; Muraki & Bock, 1991); (2) scaling the estimated parameters to a base IRT scale applying a linear transformation; (3) if true scores are used, the true scores of the new test form are converted to the true score scale of an old form. The IRT linking process in our study consists of the following steps: following the single group design, the full sample completed both assessments. Therefore, we scaled the data from both tests simultaneously in a single IRT model using the PARSCALE software. While doing this, we fixed the item parameters of the TIMSS items to values taken from the TIMSS 2011 database. Consequently, we received person parameters and item parameters for the NEPS-items on the international TIMSS metric.

In the next step, the NEPS data was calibrated separately by importing fixed item parameters from the NEPS main study 2010 in grade 5 (Duchhardt & Gerdes, 2012). This procedure resulted in person parameters based on the metric of the NEPS main study.

The linking between these two scales was conducted by using the computer software IRTEQ (Han, 2009), which allows rescaling a test form to another using various IRT scaling methods. The aim is to find a linear transformation between NEPS item parameters on the TIMSS scale and NEPS item parameters on the NEPS scale. There are several IRT scaling/equating methods that are widely used with an anchor test design (also called the “Non-Equivalent Groups Anchor Test”): mean/mean (Loyd & Hoover, 1980), mean/normal (Linn, Levine, Hastings & Wardrop, 1981), and TCC methods (Haebara, 1980; Stocking & Lord, 1983). We decided to apply only the TCC method because this method leads to the most solid results in our case.

5. Results

5.1. Conceptual overlapping of TIMSS and NEPS

As mentioned in Section 2, linking two or more studies can yield more or less strong results. This depends on the concordance of the studies. Thus, in the following test specifications, frameworks and test administration of the two studies are compared with each other. Afterwards, the results of the expert reviews are presented.

5.1.1. Content review

The first research step focuses on the conceptual equivalence between both tests. Therefore, the mathematical assessments in TIMSS and NEPS were compared in order to determine whether the same mathematical constructs were measured. Although both tests measure mathematics, some conceptual differences between the two tests’ instruments can be found:

(1) The NEPS framework is based on the mathematical literacy concept (OECD, 2003, p. 24). In comparison, TIMSS 2011 focuses on a curriculum model.
(2) Another difference between the two tests is the grade level they are created for. NEPS measures the competence of fifth graders at the beginning of the school year whereas TIMSS measures the competence of fourth graders at the end of the school year. However, this typically means a difference of approximately ten weeks in a student’s school career.

(3) Both test frameworks take different domains into account. TIMSS measures two domains: the content and the cognitive domain.2 NEPS also defines the content area, and additionally the mathematical and cognitive processes,3 but does not define the cognitive domain. The NEPS framework differentiates between four content areas (quantity, change and relationships, space and shape, data and chance), while the TIMSS framework differentiates between three content areas (number, geometric shapes and measures, data display). The areas in which the tests have a similar content are shown here with the same shades of gray (see Fig. 2). For example, “space and shape” in NEPS is conceptually almost the same as “geometric shapes and measures” in TIMSS. A difference is that “number” in TIMSS includes “quantity” and “change and relationship” in NEPS. TIMSS does not differentiate mathematical and cognitive processes as they are defined in NEPS; however, there is a correspondence between the mathematical and cognitive processes in NEPS and the cognitive domains in TIMSS. Overall, there is a high overlapping in the frameworks but the tests have a slightly different focus on the content areas.

5.1.2. Item classification

The results of the first expert review showed that there are no apparent differences between the percentage distribution of the NEPS items to the TIMSS 2011 content domain and the percentage distribution of the TIMSS items (Table 1). Only one of the NEPS items could not be classified according to the TIMSS framework as this item measures competency in probability calculation, which is not part of the TIMSS framework. TIMSS tends to have more items in the content domains “number” and “geometric shapes and measures” than the NEPS test, but a fewer number of items in “data display”.

The results of the NEPS item classification according to the TIMSS cognitive domains are shown in Table 1, too. There are no significant differences between the distribution of the NEPS items in the TIMSS framework and the distribution of TIMSS items. Generally, there are less NEPS items assigned to the TIMSS cognitive domain “knowing” than TIMSS items. In contrast, there are a few more NEPS items assigned to the TIMSS cognitive domains “applying” and “reasoning”.

The results of the second expert review also reveal a high overlap between the two tests. Regarding the formal demands, only one aspect is significantly different between the two tests (Table 2). The TIMSS test uses significantly more graphical representation like diagrams, graphics, or graphs than the NEPS test ($\chi^2 = 4.06, df = 1, p < .05$). This means the TIMSS test requires competencies in interpreting diagrams, graphics, or graphs in more items. In general, there are fewer formulas in the TIMSS test. In contrast, no tables occur in the NEPS test but in 9.1% of the items in the TIMSS test. The usage of the response format shows some

---
1 The computer software LEGS only uses positive integers. Therefore, the NEPS scores are converted by applying the transformation function $x = \text{round} \left( \frac{y}{100} \right) \times 500$.

2 The content domain describes phenomena that can be observed if the world is seen with mathematical eyes, for example numbers.

3 The cognitive domain describes the cognitive operations necessary for solving certain tasks, for example reproducing.

4 The process domain describes cognitive operations applied by students in the content areas, for example problem-solving.
Table 2 Comparison of the formal demands between the NEPS and TIMSS test.

<table>
<thead>
<tr>
<th>Item characteristics</th>
<th>NEPS</th>
<th>%</th>
<th>TIMSS</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence of formulas</td>
<td>3</td>
<td>12.5</td>
<td>18</td>
<td>10.3</td>
</tr>
<tr>
<td>Presence of tables</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>9.1</td>
</tr>
<tr>
<td>Presence of graphs</td>
<td>5</td>
<td><strong>20.8</strong></td>
<td><strong>74</strong></td>
<td><strong>42.3</strong></td>
</tr>
<tr>
<td>Presence of pictures/photos</td>
<td>5</td>
<td>20.8</td>
<td>34</td>
<td>19.4</td>
</tr>
<tr>
<td>Selected response</td>
<td>13</td>
<td>54</td>
<td>97</td>
<td>55.4</td>
</tr>
<tr>
<td>Closed constructed-response</td>
<td>11</td>
<td><strong>45.8</strong></td>
<td>52</td>
<td><strong>29.7</strong></td>
</tr>
<tr>
<td>Open constructed-response</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td>14.9</td>
</tr>
</tbody>
</table>

Bold values show statistically significant differences (p < .05).

Table 3 Comparison of the number of words, sentences, and mathematical terms between the NEPS and TIMSS test.

<table>
<thead>
<tr>
<th>Language demands</th>
<th>NEPS</th>
<th>MW</th>
<th>SD</th>
<th>TIMSS</th>
<th>MW</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of mathematical terms</td>
<td>24</td>
<td>0.96</td>
<td>2.29</td>
<td>175</td>
<td>0.97</td>
<td>1.79</td>
</tr>
<tr>
<td>Number of words</td>
<td>24</td>
<td><strong>34.67</strong></td>
<td>23.11</td>
<td>175</td>
<td><strong>23.68</strong></td>
<td><strong>15.83</strong></td>
</tr>
<tr>
<td>Number of sentences</td>
<td>24</td>
<td>2.83</td>
<td>1.52</td>
<td>175</td>
<td>2.3</td>
<td>1.45</td>
</tr>
</tbody>
</table>

Bold values show statistically significant differences (p < .05).

cumulative frequencies (cumfreq), percentage of students who scored with \( f(x) \), cumulative percentage of students who scored with \( G(y) \), and percentile rank \( Q(y) \).

Table 5 provides a selected sample of low, medium, and high NEPS scores \((x)\), converted NEPS scores \((x_c)\), frequencies \((freq)\),

Table 4 Descriptive statistics for Test Y (TIMSS).

<table>
<thead>
<tr>
<th>( y )</th>
<th>freq</th>
<th>cumfreq</th>
<th>( g(y) )</th>
<th>( G(y) )</th>
<th>( Q(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>355</td>
<td>1</td>
<td>1</td>
<td>0.001</td>
<td>0.001</td>
<td>0.008</td>
</tr>
<tr>
<td>375</td>
<td>2</td>
<td>3</td>
<td>0.003</td>
<td>0.004</td>
<td>0.273</td>
</tr>
<tr>
<td>380</td>
<td>1</td>
<td>4</td>
<td>0.001</td>
<td>0.005</td>
<td>0.478</td>
</tr>
<tr>
<td>385</td>
<td>1</td>
<td>5</td>
<td>0.001</td>
<td>0.007</td>
<td>0.615</td>
</tr>
<tr>
<td>395</td>
<td>3</td>
<td>8</td>
<td>0.004</td>
<td>0.011</td>
<td>0.888</td>
</tr>
<tr>
<td>405</td>
<td>2</td>
<td>10</td>
<td>0.003</td>
<td>0.014</td>
<td>1.230</td>
</tr>
</tbody>
</table>

Differences, too. In the NEPS test, there are no items with a constructed-response format while 14.9% of the TIMSS items have a constructed-response format. However, the TIMSS test encompasses fewer items with a short constructed-response format.

The results regarding the language demands are shown in Table 3. The NEPS test uses on average significantly more words in the mathematics items \((t = 2.26, df = 26, p < 0.5)\) and tends to have more sentences than the TIMSS test. So the NEPS test requires higher reading skills than the TIMSS test. The usage of mathematical terms is approximately the same in both tests.

5.2. Results of the equipercentile and IRT linking

The aim of this section is to find out whether the equipercentile or the IRT linking method should be preferred regarding the descriptive statistics and the classification accuracy according to the TIMSS international benchmarks.

Table 4 provides a selected sample of low, medium, and high TIMSS scores \((y)\), frequencies \((freq)\), cumulative frequencies \((cumfreq)\), percentage of students who scored with \( y \) \((g(y))\), cumulative percentage of students who scored with \( y \) \((G(y))\), and percentile rank \((Q(y))\).

Table 5 provides a selected sample of low, medium, and high NEPS scores \((x)\), converted NEPS scores \((x_c)\), frequencies \((freq)\),

Table 1 Distribution of the NEPS items in the TIMSS framework and distribution of TIMSS 2011 items referring to the content and cognitive domains.

| 2011 framework | NEPS in TIMSS | TIMSS 2011 | | | |
|----------------|---------------|------------|---|---|
| Content domains | N | % | N | % |
| Number          | 11 | 46 | 90 | 51 |
| Geometric shapes and measures | 7 | 29 | 61 | 34 |
| Data display    | 5  | 21 | 26 | 15 |
| Not classifiable| 1  | 4  |     |   |
| Total           | 24 | 100| 177| 100|
| Cognitive domains | | | | | |
| Knowing         | 7  | 29 | 71 | 40 |
| Applying        | 11 | 46 | 71 | 40 |
| Reasoning       | 6  | 25 | 35 | 20 |
| Total           | 24 | 100| 177| 100|
The results of the equating are provided in the lower part of Table 6. When applying the equipercentile equating method, the descriptive statistics show that the mean and standard deviation, but also skewness and kurtosis of the equating estimates, are very close to the TIMSS scale descriptive statistics. Therefore, equipercentile equating seems to be an adequate linking method.

In a next step, the students’ classification was based on the TIMSS 2011 international benchmarks in mathematics. The TIMSS results were then compared with the score equivalents based on equipercentile equating methods (Table 7). TIMSS 2011 (Mullis et al., 2012) reports achievement at four points along the scale as international benchmarks: Advanced International Benchmark (625), High International Benchmark (550), Intermediate International Benchmark (475), and Low International Benchmark (400). The frequency distribution of students reaching the different international benchmarks based on the TIMSS test and based on score equivalents are very close to each other. The maximum difference between TIMSS based results (9.4%) and score equivalents (10.9%) is 1.5% on the Advanced Benchmark. On a group level, this classification accuracy can be evaluated as a good approximation. Focusing on the individual level, the classification consistency of each international benchmark should also be considered. One result is that overall 54% of the students were classified as belonging to the same proficiency level. On average the classification consistency is 44%. In order to take the chance adjustment into account the kappa statistics can be calculated. The $\kappa = 0.34$ shows that the score equivalents based on the equipercentile equating method are a good approximation. However, about half of the students are classified as belonging to different proficiency levels based on the score equivalents rather than based on the TIMSS test. Therefore, the classification should not be reported or interpreted on individual level.

5.2.1. Results of the IRT scale transformation

Table 6 provides descriptive statistics of the IRT scale transformation. The transformation function estimated by IRTEQ-software for this population is $\delta_{\text{link}} = (0.93 * \delta) + 0.78$. Applying this transformation function, the transformed scores have the same mean as the non-equated TIMSS scores. However, the standard deviation is slightly different. The skewness is equal and the kurtosis is similar.

As described earlier, the students’ classification based on the TIMSS 2011 international benchmarks in mathematics was carried out. Table 7 shows that the frequency distribution of students reaching the different international benchmarks based on the...
TIMSS test and based on score equivalents are similar. The maximum difference between TIMSS based results (37.7%) and score equivalents (29.3%) arises to 8.4% on the high international benchmark. Additionally, the classification consistency was calculated. In 55% of the cases the same international benchmarks were classified. The classification consistency on average is 36%. The χ = 0.35 shows that the distribution of the individual linking is a good approximation. Also for the IRT-linking about half of the students are classified into different proficiency levels based on the score equivalents. Therefore, the classification on an individual level should not be used for interpretation.

### 5.2.2. Comparison of the outcomes of the linking methods

Table 6 provides descriptive statistics for NEPS mathematics scores, TIMSS mathematics scores, the score equivalents of the NEPS test on the metric of TIMSS based on the equipercentile linking, and the score equivalents of NEPS on the metric of TIMSS based on the IRT linking.

When we compare the results of the two equating-methods with the results of the non-equated TIMSS and NEPS test (see Table 6), we see that the transformed scores have the equivalent mean of the non-equated TIMSS scores. The standard deviation and the kurtosis of the score equivalence of the equipercentile equating are closer to the TIMSS values than to those of the IRT-linking method. Overall, the distribution seems to be reliable.

Table 7 shows the classification on a student level that seems to not be as reliable as the descriptive statistics. The score distribution on the TIMSS international benchmarks of the two linking methods compared with the non-equated TIMSS distribution is slightly different for the equipercentile linking and the IRT approach. The highest difference arises on a high international benchmark between the TIMSS (37%) and the TIMSS IRT score equivalents (29.3%). However, the Cohen's Kappa values from around 0.35 show that the distribution of the individual linking scores is sufficient in both methods. Nevertheless, to a certain extent there are differences between the proficiency level distributions. Additionally, the classification consistency was calculated. Overall, 54% of the students with the equipercentile method and 55% of the students with the IRT method were classified as belonging to the same proficiency level. The classification consistency with the equipercentile method (44%) is on average much higher than with the IRT method (36%). In comparison to other studies, the average classification consistency is good (e.g., Pietsch, Böhme, Robitzsch, & Stubbe, 2009 calculated an average classification consistency of 33% with an IRT linking approach).

### 6. Discussion

The objective of this study was to use linking methods to allow for a criterion-based interpretation of NEPS scores on the TIMSS metric. In addition, we analyzed in this context which linking method should be preferred regarding descriptive statistics and classification accuracy of the TIMSS international benchmarks. Therefore, two linking methods were compared: the equipercentile and the IRT linking approach.

#### 6.1. Conceptual overlapping of TIMSS and NEPS

With our comparison of the TIMSS and NEPS assessment frameworks we found out that the two tests are very similar in most aspects of the conceptual frameworks, but still differ to a certain extent. The tests measure mathematical competencies at the end of primary school and at the beginning of secondary school. The construct of mathematical competence is very similar. However, the two tests have a slightly different focus: the NEPS assessment is based on the concept of mathematical literacy whereas the TIMSS assessment is based on a curriculum model. Nevertheless, the comparison of the mathematical frameworks indicates a high overlapping of the content and cognitive domain. The first expert review confirms this assumption. The results of the second expert review showed that the NEPS assessment uses more words and thus calls for a higher reading proficiency in comparison to the TIMSS test. This can influence the difficulty of the test items especially for, e.g., second language learners (Haag, Heppt, Stanat, Kuhl & Pant, 2013; Wolf & Leon, 2009). On the other hand, the TIMSS items contain more diagrams, graphics, or graphs. This usage could potentially, for example, help students with reading difficulties to understand the task, as they have the visual support of the diagrams, graphics, or graphs instead of dense reading texts (Prenzel, Häußler, Rost, & Senkbeil, 2002; Schnitz, 2005). The measurement conditions are standardized in both tests, but the study design is not the same (longitudinal study vs. cross sectional study). In addition, there are differences in the methods of scaling the data (1-pl-IRT model vs. 3-pl-IRT model).

#### 6.2. Results of the equipercentile and IRT linking

The results concerning our second research question showed that the equipercentile equating method and the IRT linking method produced similar results regarding the population means and skewness. Furthermore, both methods have a satisfying classification consistency with respect to the TIMSS international benchmarks. However, regarding the estimation of the standard deviation and the kurtosis, the equipercentile method has one major advantage: namely that the distribution of the score equivalents is approximately the same. One disadvantage is the sensitivity of the linking to irregularities in the distribution of the tests. The linking, for example, is not very reliable for scores which were reached only by a few students. Thus, the equipercentile equivalents seem erratic for the very high and low achievers. One possibility to deal with this issue is pre-smoothing or post-smoothing the data (Dorans, Moses & Eignor, 2011; Livingston, 2004; Yin, Brennan & Kolen, 2004). A further limitation is the inability to estimate score equivalents for students scoring below the lowest or above the highest raw score. For example, in this study the lowest NEPS score is “145” and the highest is “900”, but in the NEPS main study some students’ scores were below “145” or
above “900”. For these students no score equivalent can be estimated. When one uses the IRT method this problem does not occur because the transformation function allows an estimation of all score equivalents. One limitation of both methods is the arising of errors through the rounding of the data. Linking the tests with IRT methods has advantages like the flexibility in linking tests. However, IRT linking is conceptually and procedurally complex, and based on strong assumptions which are hard to meet. If the flexibility of the IRT linking is unnecessary, it is better to use methods with weaker assumptions, such as equipercentile equating (Livingston, 2004).

Regarding the outcomes of the linking, it has been demonstrated that the descriptive statistics are fairly reliable, e.g., regarding mean and standard deviation. The distribution to the proficiency levels, however, is less reliable. This distribution is influenced by the reliabilities of the tests, the correlation between the tests, the parameter model used, as well as the position and the amount of cut scores (Ercikan & Julian, 2002; Pietsch et al., 2009; Stone, Weissman & Lane, 2005). Pietsch et al. (2009) provide a table showing the expected classification consistency on average for two tests linked with the IRT-method, with a given reliability of 1.00 for one test. The NEPS test has a reliability of EAP/PP reliability = .80 (Duchhardt & Gerdes, 2012) and the TIMSS test of $\alpha = .82$ (Foy, Martin, Mullis & Stancio, 2012). With a correlation of $r = .90$ the maximum expected classification consistency on average is 42%. Thus, the classification consistency on an average of 44% for the equipercentile method and of 36% for the IRT method is a good approximation. However, there are some minor differences in classification according to the international benchmarks. Thus, inferences are possible on a population level, but the results should not be reported or interpreted on an individual level.

The outcomes of this study indicate a good fit and similar results for both linking methods. This demonstrates that the linking results are reliable across the two different approaches, thus indicating that the linking has a good quality. If one was to choose one of the two linking methods, then in this case the equipercentile method should be preferred because it has weaker assumptions.

6.3. Limitations and outlook

A first limitation of the study is that it consists of a selective sample. Only schools which participated in a program for school development were selected. This school program aimed to improve teaching science and mathematics. So we must assume that the sample is selective and that the students at these schools might be better at mathematics than in schools which did not participate in this program. Secondly, we tested students at the end of fourth grade although the NEPS test is developed for students at the beginning of grade five. Hence, at least some items could be too difficult for fourth graders. Finally, there are possible order effects due to all students having taken the TIMSS test on the first day and the NEPS test on the second day. A consequent training effect or, conversely, a decrease in motivation is possible.

To give a short outlook, the aim was to find a linking function so that the NEPS scores of the main study can be interpreted in an international criterion based reference. The linking function presented here can be used to classify the results of the NEPS main study in the international TIMSS metric. For instance, the outcomes could be used as a basis for longitudinal studies on students who fail the lowest (or reach the highest) educational standards. In this manner, one could analyze to what extent the competencies which are internationally considered as crucial for fourth graders are interrelated with the educational development of students during their lifespan. In addition, factors which influence mathematical competence could be identified. Thereby, evidence based results can be used for decision-making in educational policy, e.g., for adjusting the curriculum and adopting measures to strengthen student resources.

Acknowledgements

The project was funded by the Centre for International Student Assessment (ZIB) and the Federal Ministry of Education and Research (BMBF) in Germany. The authors would like to thank the ZIB and BMBF for their support.

References


