Measuring mathematics competence in international and national large scale assessments

Ehmke, Timo; Ham, Ann-Katrin; Sälzer, Christine; Heine, Jörg; Prenzel, Manfred

Published in:
Studies in Educational Evaluation

DOI:
10.1016/j.stueduc.2020.100847

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Studies in Educational Evaluation 65 (2020) 100847

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Measuring mathematics competence in international and national large scale assessments: Linking PISA and the national educational panel study in Germany

Timo Ehmke[a,*], Ann-Kathrin van den Ham[b], Christine Sälzer[e], Jörg Heine[c], Manfred Prenzel[d]

[a] Leuphana Universität Lüneburg, Scharnhorststr. 1, D-21335 Lüneburg, Germany
[b] Leibniz-Institut für die Pädagogik der Naturwissenschaft und Mathematik, Olshausenstr. 62, D-24118 Kiel, Germany
[c] Technische Universität München, Arcisstraße 21, D-80333 München, Germany
[d] Universität Wien, Porzellangasse 4, 1090 Wien, Austria
[e] Universität Stuttgart, Azenbergstr. 16, D-70174 Stuttgart, Germany

ARTICLE INFO

Keywords:
Student evaluation
Evaluation methods
Linking study
Mathematics competencies
Mathematics achievement
Large scale assessment
Equipercentile equating
PISA
NEPS

ABSTRACT

This study examines a linkage between the international mathematics scale of the Programme for International Student Assessment (PISA 2012) and the mathematics assessment taken from the German National Educational Panel Study (NEPS). The linking was realized by a separate linking study that uses a single group design. The sample consists of $n = 1{,}270$ 9th graders from 78 German secondary schools. The equipercentile linking leads to close descriptive scale characteristics (means, standard deviation, and skewness) between the original PISA mathematics scale and the PISA score equivalents. The linking was stable over the four subgroups (gender, migration background, books at home, and school type). Altogether, the results indicate that assigning students to PISA proficiency levels given their NEPS mathematics test score, the PISA score equivalents produce a similar distribution of students reaching the PISA proficiency levels at a group level.

1. Introduction / Background

Mathematical competence can be seen as an important prerequisite for lifelong learning and active participation in society and culture. Therefore, the mathematical competence of students from different countries is regularly measured by strategies such as the Programme for International Student Assessment (PISA) and national large-scale assessments such as the National Educational Panel Study (NEPS) in Germany. Although these assessments seem to measure a similar characteristic (e.g., mathematics competence), comparing the results from NEPS with the results reported in OECD's PISA 2012 is not feasible, just as comparing temperatures in different capital cities would not be possible without knowing whether the Fahrenheit or Celsius scale is being used (Cartwright, Lalancette, Mussio, & Xing, 2003). However, is it possible to link the mathematics measurement scales from NEPS and PISA? To answer this question, this study analyzes the linking of the two measurement scales and evaluates the robustness of this link. A linkage between mathematics tests from NEPS and PISA 2012 could enable educators and researchers to report and interpret the results of NEPS's mathematical assessment in relation to PISA's international benchmarks, which would be especially helpful for longitudinal analyses of at-risk students and an extension of the research possibilities of NEPS.

Current state of research in linking studies including PISA or NEPS

To date, there have been no studies that connect the respective measurement scales of PISA and NEPS. However, with increasing frequency, there have been attempts to link international studies such as PISA with other large scale assessments. To provide an overview of linking studies where either PISA or NEPS is involved, we grouped the studies regarding three different research purposes (Ehmke, Köller, van den Ham, & Nissen, 2014; van den Ham, Ehmke, Nissen, & Roppelt, 2017):

(1) Comparing studies to explain the differences in outcomes

One research purpose is to show the differences and commonalities between multiple large-scale assessments (LSA) so that the results of the studies can be better interpreted and differences in outcomes can be explained. For instance, Neidorf, Binkley, Gattis, and Nohara (2006) compared the mathematics frameworks and items included in the

* Corresponding author.
E-mail address: tehmke@leuphana.de (T. Ehmke).

https://doi.org/10.1016/j.stueduc.2020.100847
Received 19 May 2017; Received in revised form 3 February 2020; Accepted 5 February 2020

National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and PISA 2003. They found that NAEP has more similarities with TIMSS than with PISA, and they recorded the studies' main differences. They stated that more in-depth analyses of the items would reveal more important differences. Overall, they concluded that the three assessments are complementary. Therefore, for one specific topic or skill, one study might provide more information 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 with a regression analysis. One of her conclusions was that content balance and sampling definitions affect the outcomes at the country level. Furthermore, the PISA assessment instruments used more words than the TIMSS' tests, and students with lower reading competencies performed unequally in TIMSS and PISA.

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

Another research purpose is to compare one or more studies in relation to differences in the national and, in the case of PISA, international benchmarks. Thus, it can be verified whether the requirements in the national study are higher or lower than its international counterpart. This was the research purpose of Hambleton, Sireci and Smith (2009). They compared NAEP with the TIMSS and the PISA study in 2003. The research question was whether the proficiency levels in NAEP are set too high, but the results showed that this was not the case in the international context.

Cartwright (2012) established a connection that links the British Columbia English Examination and the PISA Reading Scale. The results of an initial review and comparison of the assessment frameworks indicate that the two assessments measure a similar general construct of reading. After the application of a linking function (kernel equating), the equivalences between the proficiency scale levels on the BC English reading and PISA reading scales could be compared. This linkage allows for the interpretation of the BC Exam scores on the international PISA reading measurement scale. For example, what does an "A" level performance in British Columbia mean in the international PISA context (Cartwright, 2012)?

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

Linking a national test to an international test provides the opportunity to locate the outcomes of the national study on the international scale. As a consequence, it is possible to, for example, use the international benchmarks and, with them, a reference criterion 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 findings 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.

Nissen, Ehmke, Köller and Duchhardt (2015) linked the mathematics measurement scale of the National Educational Panel Study (NEPS, 4th graders) with the criterion-based reference frame of the TIMSS 2011. The results indicate a high conceptual overlap between NEPS and TIMSS. To link the studies, two different methods—the equipercentile linking and an Item response theory (IRT) linking approach—were compared to distinguish their descriptive statistics and classification accuracy from those designated by TIMSS' international benchmarks. The main result of the linking process (equipercentile linking and an IRT linking approach) was that both methods showed similar descriptive statistics and a satisfactory classification consistency.

Wagner, Hahn, Schöps, Ihme and Köller (2018) evaluated an equipercentile equating between the science scale of the German National Educational Panel Study (NEPS) and the science scale of the Programme for International Student Assessment (PISA 2012). The linking via equipercentile equating showed that the linking function between the NEPS and PISA science scores can be used to classify the NEPS science scores within the international science benchmarks of PISA.

1.1. The study designs of NEPS and PISA

In Germany, a NEPS was established in 2010 to provide a large database for analyzing the educational processes of large samples throughout the entire lifespan. The cohorts who participate in the study range from newborns to 65-year-old adults. A large sample of a cohort of ninth graders started in 2010. The longitudinal NEPS complements the cross-sectional LSA in Germany. For instance, since 2000, Germany has routinely participated in the PISA. The following subsections provide detailed information about the study designs of PISA and NEPS and ultimately compare the assessment frameworks.

1.1.1. About PISA

In the year 2000, the first *Programme for International Student Assessment* (OECD, 2001) was conducted by the *Organisation for Economic Co-operation and Development* (OECD). PISA 2012 is the fifth cycle of a cross-sectional study that assesses the reading, mathematics, and science competencies of fifteen-year-old students. As it is conducted every three years, PISA provides the opportunity to measure trends for a collection of age cohorts in the participating educational systems. The overall aim of this study is to compare the outputs of countries' educational system through regular monitoring. PISA can serve to create benchmarks between countries and helps to identify the factors that could influence the various countries' differing results (Seidel & Prenzel, 2008). In 2012, 65 countries participated in PISA (OECD, 2013). In addition to the assessment instruments in mathematics, science, and reading, students' social background and other characteristics are collected via questionnaires. The basis for the tests' frameworks is formed by approaches to theoretical models of literacy (OECD, 2013).

1.1.1.1. PISA 2012 – Mathematics test for fifteen-year-old students. The PISA 2012 assessment for fifteen-year-old students takes a total of 120 min for the cognitive test and 30 − 35 min for the student questionnaire. Participation in the test was obligatory for the selected sample in Germany. The cognitive test contains 110 mathematics items in addition to reading and science items. All items are distributed across 13 booklets, and each composes four 30-minute blocks in total. Between one and three out of four blocks contain mathematics tasks. The test comprises three item response formats, namely, an open constructed-response, closed constructed-response and selected-response (multiple-choice). The majority of the items are scored dichotomously. The open constructed-response items can sometimes involve partial credit scoring. IRT is used to scale the data and develop the reporting scales. The students' ability scores are estimated by weighted likelihood estimates (WLE – Warm, 1989) and, for the final international reports, by the plausible values technique (OECD, 2014). PISA also reports the proficiency scale levels in three domains (mathematics, science, and reading). In mathematics, for example, six proficiency scale levels are defined that describe different levels of students' mathematical capabilities.

PISA 2012 defines mathematical literacy in terms of the following three interrelated aspects (OECD, 2013, p. 27): (A) the mathematical processes that describe what individuals do to connect the context of the problem to mathematics to thus solve the problem, and the capabilities that underlie these processes; (B) the mathematical content that is targeted for use in the assessment items; and (C) the contexts in which the assessment items are located. The mathematical processes (A) can be structured in the following three categories: (1) formulating

situations mathematically; (2) employing mathematical concepts, facts, procedures, and reasoning; and (3) interpreting, applying and evaluating mathematical outcomes. Seven mathematical capabilities underpin each of the processes as follows: (1) communication; (2) mathematizing; (3) representation; (4) reasoning and argument; (5) devising strategies for solving problems; (6) using symbolic, formal and technical language and operations; and (7) using mathematical tools. The mathematical content (B) consists of four categories, namely, (1) change and relationships, (2) space and shape, (3) quantity, and (4) uncertainty and data. The used contexts (C) are classified into the four categories of (1) personal, (2) occupational, (3) societal, and (4) scientific. The items evenly cover the mathematical content categories and contexts. Approximately 50 % of the items can be allocated to the second process, 25 % can be allocated to the first process and 25 % can be allocated to the third process (OECD, 2013).

1.1.2. About NEPS

As stated above, the National Educational Panel Study (NEPS) is a longitudinal study in Germany. Its aim is to measure the evolution of various competencies over a subject's lifespan (from early childhood to late adulthood), and educational processes and trajectories (Blossfeld, von Maurice, & Schneider, 2011). In addition, the study measures the development of competencies and the impact of learning opportunities. Therefore, NEPS uses a multicohort-sequence design. There are six starting cohorts at different stages of transition in the educational system: newborns, four-year-old students, fifth graders, ninth graders, university freshmen and higher education students and adults aged 25–65 years old. Four starting cohorts began in 2010, one began in 2009, and one began in 2012. The participants are to be tracked over the course of their lifespans.

NEPS was initiated and is financed by the German Federal Ministry of Education and Research (Blossfeld, Doll, & Schneider, 2009). Since 2014, NEPS has been organized and coordinated by the *Leibniz Institute for Educational Trajectories* (LIfBi). NEPS assesses the following four domains (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.

1.1.2.1. NEPS 2010, grade 9 – mathematics test. In this study, we used the NEPS mathematics test for ninth graders (NEPS-K9) at the beginning of their school year. Participation was optional. The NEPS mathematics test is based on a theoretical model of mathematical literacy (Weinert et al., 2011) that is inspired by and similar to the PISA framework (OECD, 2010). The NEPS framework differentiates between two aspects: the content areas and the mathematical and cognitive processes. The content areas include the subdomains of quantity (33 %), change and relationship (25 %), space and shape (21 %), and data and chance (21 %). These subdomains correspond to the subdomains used in PISA 2012. The mathematical and cognitive processes differentiate among argumentation, communication, mathematical modeling, mathematical problem solving, representing, and applying technical skills (Weinert et al., 2011).

The test for ninth graders takes 112 min overall. In 2010, the assessment used two different booklets, and each contained 22 mathematics items, 28 science items, 51 reading items, 40 information and communications technology (ICT) literacy items, 89 listening comprehension items, and 4 domain-specific procedural metacognition items. The position of the ICT and science domains rotated in the two booklets, but the mathematics domain had the same position in both booklets. In NEPS, the Rasch model is used to calculate WLE scores as estimates for the students' achievement scores (Duchhardt & Gerdes, 2013). However, proficiency scale levels are not defined in NEPS. The mathematics items are multiple choice (19 items), complex multiple choice (2 items), or constructed response (1 item). Twenty items are dichotomous, and two are partial credit. No multimatrix design was applied regarding the choice and order of the items within the mathematics test. All students received the same mathematics items in the same order.

1.1.3. Comparison of the mathematics assessment frameworks between PISA and NEPS

As a precondition to establishing a statistical linkage, the first step of the process involved a comparison and review of the assessment frameworks used in both assessments. Therefore, a detailed review was conducted by content specialists with expert knowledge of both the mathematics frameworks and items from the NEPS assessments and the PISA 2012 study. van den Ham, Nissen, Ehmke, Sälzer and Roppelt (2014) published the detailed results of this conceptual comparison. According to this work, experts reviewed the mathematics items of the PISA 2012 and NEPS-K9 studies regarding their distribution among subdimensions, their formal and language demands, and their linguistic complexity. The results show that the tests are very similar in terms of the frameworks. One finding concerning the comparison of the PISA 2012 and NEPS-K9 items reveals that PISA 2012 mathematics items are more difficult regarding the word level and the complexity of the sentence structure. However, the salient finding of this review was that overall, both assessment frameworks show a high conceptual overlap, which seems to demonstrate a link between the two assessments. The review provides evidence that both assessments are based on similar constructs and that a statistical linking between both scales is feasible.

Such an interpretation of linked scores is bound to several preconditions. In addition to the high conceptual overlap (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Kolen & Brennan, 2010; Linn, 1993; Mislevy, 1992) the following requirements are specified in the literature: (1) high correlations between the tests that are to be linked, and (2) a sufficient distributional similarity between the tests are a necessity (Dorans, 2004; Hanson, Harris, Pommerich, Sconing, & Yi, 2001). Furthermore, (3) the degree to which a linkage varies across different subpopulations should be evaluated to ensure that the linkage function is appropriate for these subgroups. Moreover, (4) the classification based on linked scores has to be sufficient. By evaluating the classification consistency, it can be examined the extent to which the classification of the students is consistent regardless of which test is used (Dorans, 2004; Hanson et al., 2001; Kolen & Brennan, 2010; Pommerich, Hanson, Harris, & Sconing, 2004).

Accordingly, this study establishes the linkage of the NEPS and PISA mathematics tests and evaluates the correlation and the distributional similarity between the two tests, the stability of the linkage across subgroups and the classification consistency (Dorans, 2004; Hanson et al., 2001; Kolen & Brennan, 2010; Pommerich et al., 2004). For the linking we apply the equipercentile linking method (see section 3.3).

2. Research questions

To examine the possibility of reporting and interpreting the results of NEPS's mathematical assessment in relation to PISA's international measurement scale based on a linkage between the mathematics tests, this study analyzes the extent to which this linking produces equivalent test scores. We refer to the linked scores in the following questions as "PISA score equivalents." The research questions are:

(1) To what extent are the mathematics score distributions of PISA 2012 and of NEPS-K9 comparable, and how are the two tests correlated?
(2) Which linking function can be found by applying the equipercentile linking method to the mathematics scales of PISA 2012 and NEPS-K9?
(3) Is the derived linking function invariant across different subgroups?
(4) How accurate is the assignment of students to the PISA proficiency levels based on the original PISA scores and on the PISA score

equivalents?

3. Method

3.1. Data collection

The international PISA 2012 study and the German NEPS are two separate studies that share neither common test items nor common student populations (age-based vs. grade-based sampling). However, to link these two large scale assessments, a separate linking study funded by the Centre for International Student Assessment (ZIB) and the Federal Ministry of Education and Research (BMBF, funding reference: LSA009) in Germany was conducted by using a single group design.

This study was arranged as a two-day assessment at schools that participated in a program for school development. On the first day, every student completed one PISA test booklet, and on the second day, the students completed the NEPS mathematics tests for 9th graders. In this linking study, we used only five out of thirteen PISA booklets. We focused on the five booklets with the most mathematics (and science) items. The single group design allows for linking the scales between the PISA and NEPS mathematics assessments.

The test booklets used in the linking study were the same as the test booklets used in the main studies of PISA 2012 and NEPS starting cohort 4, grade level 9 (NEPS-K9). Therefore, it is possible to link the assessment scales of (A) the PISA test of the linking study to the international PISA test of the PISA 2012 study and (B) the NEPS test of the linking study to the NEPS test of the NEPS 2010 study (see Fig. 1). Due to the evaluation of different student populations with the same test instruments, the data collection design can be classified as being nonequivalent groups with an anchor test design (NEAT; von Davier, Holland, & Thayer, 2004). The linking study provides the opportunity to (C) link the NEPS and the PISA tests of the linking study because the same students took both mathematics tests (single group design), which is the aim of this study. With this design, it is also possible to transfer the linking function established in step C to the main studies: PISA 2012 and NEPS 2010 (D).

The data set realized in the linking study consists of 1,678 9th graders who took the PISA mathematics test and 1,330 ninth graders who took the NEPS mathematics test from 78 German secondary schools with 80 classes. On average, the students were 15 years old. Overall, n = 1,270 ninth graders 50 % male, 50 % female) took both mathematics tests from NEPS and from PISA. Almost half of the students out of the sample follow the academic track (45 %). The sample was drawn from within five German federal states stratified by school type. The participation rate of the sampled classes was approximately 79 %. The study was conducted in the spring/summer of 2012.

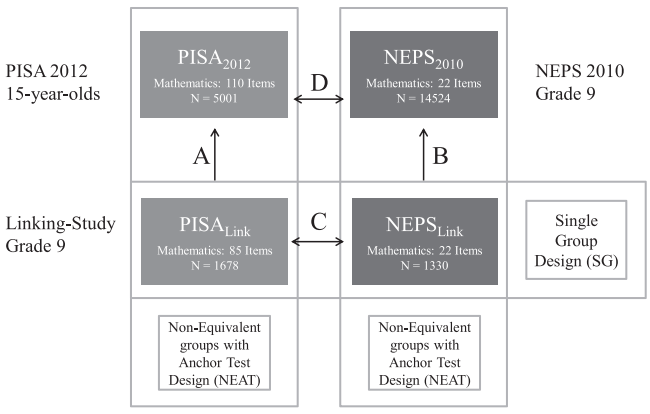

Fig. 1. Data collection design.

3.2. Scoring and data procedures

We scored the PISA 2012 mathematics data in the linking study by applying the international coding rules from PISA 2012 (OECD, 2014, technical report). The test scores were computed by using a Rasch model with fixed item parameters from the international database of PISA 2012 (OECD, 2014, technical report). Students' WLE scores were then transformed into the international PISA 2012 achievement scale metric with the transformation equation provided in the PISA technical report (OECD, 2014).

We scored the 22 mathematics items from the grade nine assessment by using the NEPS' coding rules. The raw scores were scaled by a Rasch model with fixed item parameters taken from the NEPS' 2010 main study (Duchhardt & Gerdes, 2013). The WLE person separation reliability was 0.796 for the NEPS test and was 0.791 for the PISA 2012 assessment.

3.3. Linking procedures

To link the mathematics scales from PISA 2012 and NEPS 2010, different linking approaches are available. Linn, McLaughlin, and Thissen (2009) provide an overview on three broad categories of linkages by categorizing them into the contexts in which each occurs and providing examples of the questions with their answers (see also the overview in Dorans, Pommerich, & Holland, 2007; Holland, 2007). These linkage categories are (1) equating two tests X and Y, (2) aligning the scales of X and Y, and (3) predicting Y from X. For our purposes, the second approach, namely, aligning the scales of two tests, is most suitable. Here, two different approaches are possible (Linn, McLaughlin, & Thissen, 2009): calibration or vertical scaling (often IRT-based methods) and concordance (equating-like methods).

Calibration or vertical scaling 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 the following three steps (Kolen & Brennan, 2010): (1) estimating the item parameters (e.g., using PARSCALE), (2) scaling the estimated parameters to a base IRT scale by applying a linear transformation, and (3) if true scores are used, converting the true scores of the new test form into the true score scale of an old form. IRT methods have some advantages, such as flexibility in choosing a linking plan or adaptive testing, but they are conceptually and procedurally complex. They are based on strong assumptions that often do not hold for the testing situation. If the flexibility of IRT methods is not required, it is recommended to use simpler methods with less strong assumptions (Livingston, 2004).

The concordance approach (Pommerich, 2007) uses equating-like methods (e.g., equipercentile linking). The creation of a concordance uses the statistical methods of equating to match the scores on tests that do not meet the requirements for equating. The context makes the difference; for concordance, the tests are constructed with similar but not identical frameworks and specifications (Linn et al., 2009). A concrete statistical method for linking two measurement scales and calculating a concordance table is *equipercentile linking*.

In our study, we decided to apply the equipercentile linking approach for three reasons.

First, Linn et al. (2009) recommend IRT-based approaches for measures of the same construct but with different levels of reliability, and/or difficulty. For tests that measure similar but not identical constructs, they suggest equating methods such as equipercentile linking.

Second, the equipercentile equation approach is robust against differences in the distributions of the two tests, as is the case in our study (see section 4.1). Utilizing equipercentile equating facilitates "the distribution of two forms to differ from one another in all four of the primary statistical moments" (Holmes Finch & French, 2019, p. 358).

Third, the studies from Cartwright (2012) and Nissen et al. (2015) compared the outcomes for IRT and equipercentile linking. Both studies

conclude that the equipercentile method leads to a better representation of the distributional characteristics (standard deviation, skewness, and kurtosis) between the score equivalents and the original scale compared to the IRT equating method.

The equipercentile method uses a linking curve (respectively a linking function) to depict the form-to-form differences in difficulty between two tests (Kolen & Brennan, 2010). According to the equipercentile method, the first step is to determine percentile rankings for the score distributions. After this, the scores of the two tests with the same percentile rank are declared to be equivalent (Kolen & Brennan, 2010; Muraki, Hobo, & Lee, 2000). In contrast to the mean and linear method, this method has the desirable property of always being within the range of possible scores (Kolen & Brennan, 2010). All of the individual steps for applying the equipercentile linking are described in the results section.

The equipercentile linking in our study was conducted with the computer software LEGS 2.01 (Brennan, 2004) that allows linking with equivalent or single group designs. As data input for the LEGS linking, we used the frequency score tables for the full sample and for the subsamples, as provided in Tables 2 and 3. The computer software LEGS technically only uses positive integers. Therefore, the NEPS scores are converted by applying the transformation function $x_t = \text{rnd}(x \cdot 100) + 500$. The equipercentile linking was performed with LEGS data smoothing by using the cubic-spline smoothing algorithm (Brennan, 2004). In accordance with the conventions established by Brennan (2004), the smoothing factor slim = 0.5 was chosen. The scale score range was chosen to reach from the lowest to the highest scores of the assessment scales with no truncation.

4. Results

In this section, the results of our analyses are presented according to the research questions derived above.

4.1. Comparability of the score distributions of PISA 2012 and NEPS-K9

Prior to linking the two studies, the distributional characteristics of the PISA test scores and NEPS test scores from the linking study are compared to answer the first research question. Therefore, the distributions are tested for normality and are compared to one another. Table 1 compares the distributional characteristics of the linking study of PISA 2012 to the distributional characteristics of the NEPS-K9 mathematics tests. The mean and standard deviations are not directly comparable due to the different reporting scales. However, the NEPS scores are significantly platykurtic and positively skewed compared to the distribution of the PISA scores.

The covariation of the two tests is shown in a scatterplot (Fig. 2) and quantified through the calculation of the manifest and latent correlations. Accordingly, both tests are scaled in a two-dimensional model by using the ConQuest software (Adams, Wu, & Wilson, 2012).

The covariation between the students' mathematics scores in PISA and NEPS indicates a moderate relationship between the two assessments. The observed correlation is $r = .68$ (Fig. 2), and the latent correlation as a result of two-dimensional IRT scaling is approximately $r = .90$.

Referring to the results of the initial review and the comparison of the mathematics assessment frameworks indicates that the two assessments measure a similar construct of mathematical achievement.

Table 1
Descriptive statistics for NEPS and PISA mathematics scores.

	N	Min	Max	MW	SD	Skewness	Kurtosis
NEPS	1270	−3.47	3.75	0.53	1.27	0.15	−0.42
PISA	1270	280	809	541	78	−0.02	0.09

Table 2
PISA mathematics score frequencies, percentages, and percentile ranks.

PISA score, y	freq	cum freq	f(y)	F(y)	P(y)
280	1	1	0.001	0.001	0.039
284	1	2	0.001	0.002	0.118
313	1	3	0.001	0.002	0.197
326	3	6	0.002	0.005	0.354
341	2	8	0.002	0.006	0.551
342	2	10	0.002	0.008	0.709
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
529	14	575	0.011	0.453	44.724
533	14	589	0.011	0.464	45.827
535	12	601	0.009	0.473	46.850
537	16	617	0.013	0.486	47.953
539	13	630	0.010	0.496	49.094
543	23	653	0.018	0.514	50.512
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
752	1	1264	0.001	0.995	99.488
761	2	1266	0.002	0.997	99.606
767	1	1267	0.001	0.998	99.724
771	1	1268	0.001	0.998	99.803
793	1	1269	0.001	0.999	99.882
809	1	1270	0.001	1.000	99.961

Table 3
NEPS mathematics score frequencies, percentages, percentile ranks, and PISA score equivalents.

NEPS score, x	x_t	freq	cumfreq	*f(x)*	*F(x)*	*P(x)*	PISA score equivalents
−3.47	153	1	1	0.001	0.001	0.039	279.756
−2.74	226	2	3	0.002	0.002	0.157	317.198
−2.55	245	1	4	0.001	0.003	0.276	326.944
−2.49	251	1	5	0.001	0.004	0.354	330.021
−2.48	252	1	6	0.001	0.005	0.433	330.534
−2.26	274	4	10	0.003	0.008	0.630	341.818
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
−0.01	499	28	474	0.022	0.373	36.220	516.145
0.02	502	1	475	0.001	0.374	37.362	518.044
0.03	503	1	476	0.001	0.375	37.441	518.189
0.05	505	8	484	0.006	0.381	37.795	519.885
0.07	507	1	485	0.001	0.382	38.150	521.568
0.08	508	1	486	0.001	0.383	38.228	521.623
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
2.27	727	19	1145	0.015	0.902	89.409	635.281
2.42	742	4	1149	0.003	0.905	90.315	637.058
2.53	753	61	1210	0.048	0.953	92.874	650.574
2.84	784	8	1218	0.006	0.959	95.591	681.974
3.22	822	49	1267	0.039	0.998	97.835	690.474
3.75	875	3	1270	0.002	1.000	99.882	808.388

Additionally, the assumption of a sufficient correlation between the tests can be validated. However, the equity among the score distributions is insufficient to support the notion of interchangeability between the PISA and NEPS scores.

4.2. Determination of the linking functions

The second research question concerns the determination of the linking function. The idea behind the equipercentile method is to find the percentile for a particular score on one test (PISA 2012) and equating it to the score on the other test (NEPS 2010) that is at the same percentile (Holmes Finch & French, 2019, p- 358). Table 2 provides a selected sample of low, medium, and high PISA mathematics scores (*y*),

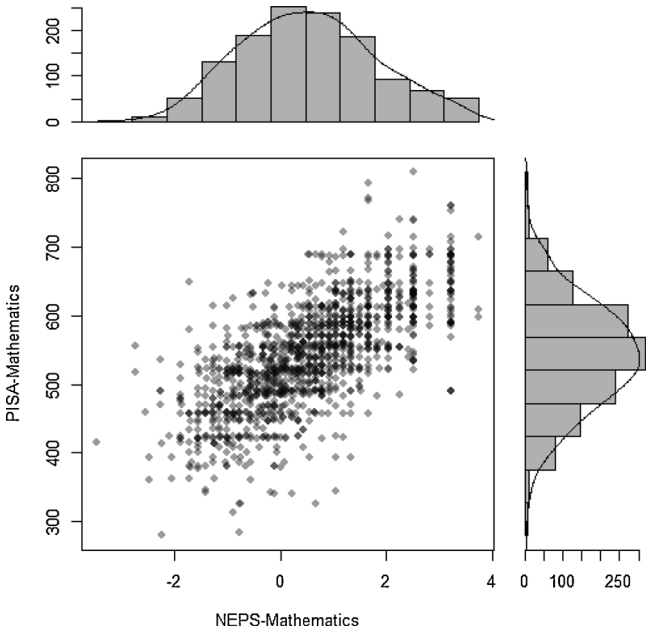

Fig. 2. Scatterplot of the PISA mathematics scores and the NEPS mathematics scores.

frequencies, cumulative frequencies, percentage of students who attained a certain score ($f(y)$), cumulative percentage of students who reached a certain score ($F(y)$), and the percentile rank ($P(y)$).

The outcomes of the equipercentile linking for a selected sample of low, medium, and high NEPS WLE scores (x), frequencies, cumulative frequencies, percentage of students who reached a certain WLE score ($f(x)$), cumulative percentage of students who reached a certain WLE score ($F(x)$), and the percentile rank of the WLE score ($P(x)$) are presented in Table 3. The last column in Table 3 contains the PISA score equivalents as a result of the equipercentile method. The PISA score equivalents are score values in the PISA measurement scale. They are based on the NEPS scores that have been transformed into the PISA measurement scale through equipercentile linking.

4.3. Invariance of linking across subgroups

The third research question refers to the robustness of the linkages across different subgroups. First, we compared the linking curves for the subsamples of boys and girls. Each of these subsamples is represented by one curve in Fig. 3. Both linking methods can now be described by the overlap – or the discrepancy – between the curves for

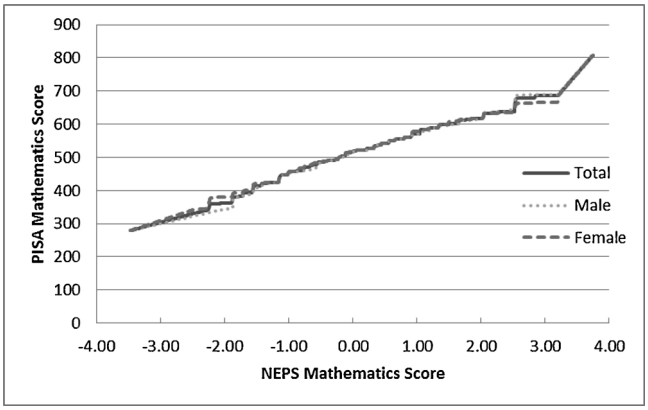

Fig. 3. Matching of NEPS scores and the score equivalents on the PISA 2012 mathematics scale.

the total sample, on the one hand, and the curve for the male or female subsample, on the other hand. As seen in Fig. 3, the curves show a high degree of overlap and, therefore, only small discrepancies for both subgroups. A closer observation of the discrepancies shows that the divergences in the equipercentile linking scores appear primarily in the lower (less than -1.8 NEPS score values) and upper score ranges (more than 2.7 of the NEPS score values) rather than in the center of the distribution. However, these discrepancies appear only for a small part of the sample (less than 7 %). They can also be explained by the fact that the error of the equipercentile linking function increases according to the inverse of the product of the two population densities. There are only a few students with very high or very low scores. This limitation in the sample results is a linking function that has limited exactness for these score sections.

The descriptive statistics of the PISA mathematics scores and the PISA score equivalents are compared in Table 4 for the total sample and for the following four different subgroups: (1) male vs. female, (2) students with a migration background vs. students without a migration background, (3) more than 100 books at home vs. less than 100 books at home, and (4) school type = grammar school (in Germany: Gymnasium) vs. another type of school.

The linking results show that equipercentile linking leads to means and standard deviations of the PISA score equivalents that are very close to the PISA 2012 scores. However, the score distributions are slightly different in skewness and kurtosis. The analysis for the four subgroups provides evidence that the linking methods are invariant to the population used in conducting the linking (Huggins & Penfield, 2012). Within each of the four subgroups, the descriptive statistics (mean and standard deviation) of the PISA score equivalents are close to the PISA 2012 scores. The results show that the linking method is stable concerning gender, migration, social background (books at home), and school type. Violations of the population invariance in this linking would have jeopardized the fairness and validity of the test scores, but this result could not be observed in this linking study. Overall, the equipercentile linking method reproduces the distribution characteristics quite well.

4.4. Classification consistency concerning the PISA proficiency score levels

The last research question refers to the classification consistency concerning PISA proficiency levels based on the PISA 2012 scores and on the PISA score equivalents. Cross-classification of the proficiency levels of the two assessments is a common approach in linking studies (Linn et al., 2009). After linking the two scales, the cut scores that define the PISA 2012 mathematics proficiency levels were applied to the PISA score equivalents. NEPS scores and PISA proficiency levels are compared in Fig. 4.

To compare the classification of students on the PISA 2012 proficiency levels based on the PISA test scores to the classification based on the PISA score equivalents, both distributions for the total sample and the examples for the male and female samples are compared. The results of the students' distribution on the PISA 2012 mathematics proficiency levels for the total sample are given in Fig. 5, and the results for the gender subgroups are presented in Fig. 6 (male) and Fig. 7 (female). Comparing these results to the frequency distribution of the students who reach the different proficiency levels based on the PISA score equivalents shows that both measures have very similar distributions patterns. Only slight differences are provided in Figs. 5–7. The maximum difference in the percentages of the PISA proficiency levels based on the PISA test results and PISA score equivalents is less than 2 % at Level 3 (29.8 vs. 27.9 %). For the two subpopulations, these differences are slightly higher.

Analyzing the classification consistency at the individual level in contrast to the group level, as shown in Figs. 5–7, reveals that approximately 42 % of the students from the total sample are classified to the exact same proficiency score level, regardless of whether the PISA

Table 4
Comparing descriptive statistics between PISA 2012 mathematics scores and PISA score equivalents.

		N	Min	Max	MW	SD	Skewness	Kurtosis
PISA 2012 Mathematics	Total	1270	280	809	541	78	−0.02	3.09
	Male	637	313	809	553	79	−0.15	3.31
	Female	633	280	761	529	74	0.06	2.92
	Migration background	126	326	690	515	71	0.01	2.81
	No Migration background	1144	280	809	544	78	−0.04	3.12
	More than 100 books at home	656	280	809	556	76	−0.01	3.20
	Less than 100 books at home	614	284	761	526	77	−0.03	2.97
	School type = grammar school	680	313	809	566	75	−0.12	3.41
	School type = other	590	280	716	513	71	−0.01	2.97
PISA Score Equivalents	Total	1270	280	808	541	76	−0.09	2.96
	Male	637	280	808	553	78	−0.24	2.96
	Female	633	280	808	529	74	0.10	3.20
	Migration background	126	280	808	516	73	0.30	3.62
	No Migration background	1144	280	808	543	76	−0.14	2.92
	More than 100 books at home	656	280	808	555	74	−0.17	2.89
	Less than 100 books at home	614	280	808	526	77	0.02	3.13
	School type = grammar school	680	280	808	565	73	−0.24	3.28
	School type = other	590	280	808	513	71	−0.02	2.85

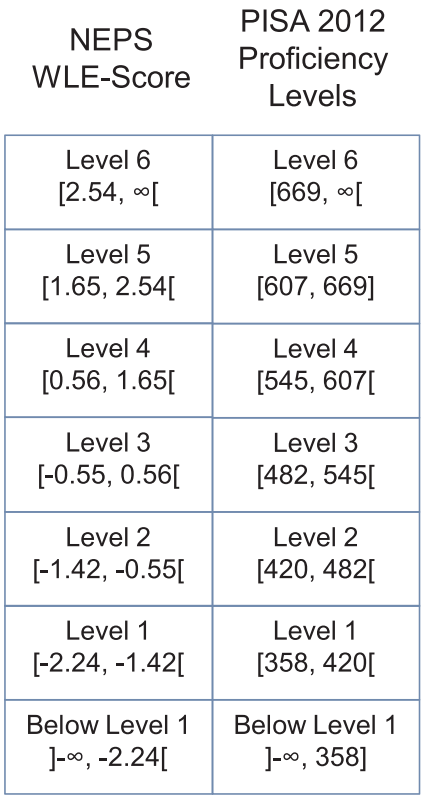

Fig. 4. Mapping between NEPS WLE scores and PISA 2012 proficiency score levels.

mathematics score or the PISA score equivalents were used (Table 5). Cohen's Kappa statistic is approximately $\kappa = .26$, which can be judged as a fair classification consistency (Landis & Koch, 1977). For the subpopulations of male or female students, the classification consistency is quite similar. For the female sample, 42.3 % of the students are classified as being on the same proficiency level, and Cohen's Kappa statistic is approximately $\kappa = .25$. For the male sample, 41.6 % of the students are classified as being on the same proficiency level, and Cohen's Kappa statistic is approximately $\kappa = .25$. The Spearman rank-order correlation coefficient for the full sample is $r_S = .68$ (female sample: $r_S = .67$, male sample: $r_S = .65$).

Altogether, the results indicate that when assigning students to proficiency levels according to their test score, the PISA score equivalents produce a very similar distribution of students who reach the PISA proficiency level at a group level compared to the frequency distribution based on the original PISA mathematics items. However, the classification accuracy at the individual level is much less reliable.

5. Discussion

The research purpose of this study was to evaluate the concordance between the mathematics tests of NEPS-K9 and PISA 2012 that could enable educators and researchers to report and interpret the results of NEPS's mathematical assessment in relation to PISA's international benchmarks. The assumptions for such an interpretation of linked scores are a (a) high conceptual overlap (Feuer et al., 1999; Kolen & Brennan, 2010; Linn, 1993; Mislevy, 1992), (b) high correlation between the tests that are to be linked, (c) sufficient distributional similarity (Dorans, 2004; Hanson et al., 2001), (d) invariance of the linked scores across relevant subgroups, and (e) consistency of classification based on the linked scores (Dorans, 2004; Hanson et al., 2001; Kolen & Brennan, 2010; Pommerich et al., 2004). Until now, only the conceptual overlap between the NEPS and PISA mathematics tests has been analyzed (van den Ham et al., 2014). Therefore, this study addressed the (1) comparability of the score distributions and correlation of the PISA 2012 and NEPS-K9 mathematics scores, (2) the establishment of a linking function by applying an equipercentile linking method, (3) the invariance of the linking across subgroups, and (4) the classification consistency to the PISA 2012 proficiency score levels based on PISA score equivalents.

Regarding the first research question, we find that there are some differences that remain in the distributional characteristics between the two score distributions. Although both assessments share a high conceptual overlap, both scales cannot be seen as identical measures. However, the correlation between the two tests is substantial (manifest: $r = .68$, latent: $r = .90$). This correlation can be seen as a necessary prerequisite for the linking approach that is pursued in the present study.

For the second research question, the linkage was established with the equipercentile method. The score distribution of the PISA score equivalents produced by the equipercentile linking is very close to the PISA score distribution.

The third research question focused on the invariance of the linking function across several subgroups. The results showed that the linking resulted in similar means, standard deviations and skewness for the four subgroups (gender, migration background, books at home, and school type). These results provide evidence for the robustness of the linking between both assessments.

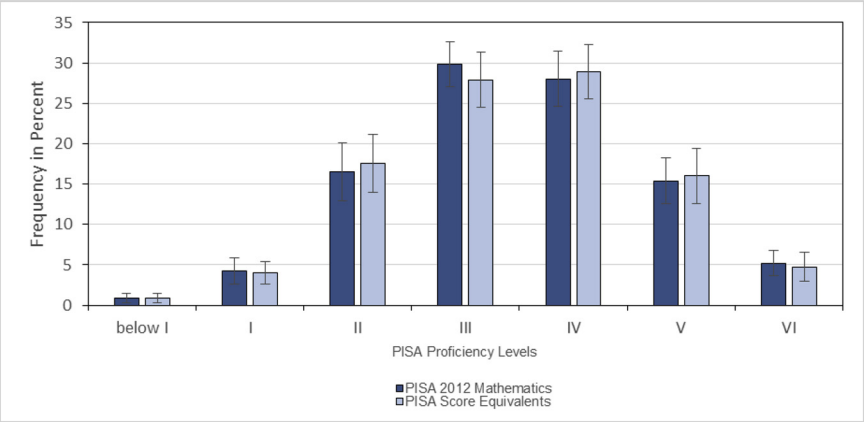

Fig. 5. Classification of students (full sample) to PISA 2012 proficiency levels based on the PISA 2012 mathematics scores and the PISA score equivalents.

From a theoretical perspective, the degree of exactness of an equipercentile linking depends on the following three parameters: (1) the number of discrete scale score values in both distributions, (2) the density of observed scores in the desired score range, and (3) the technical parameters of tuning the linking function. In practical applications, as in the present study, the number of discrete scores is more or less determined by the number of possible scores, as represented by the WLE scores that serve as an estimator for the PISA students' mathematics competence. In this respect, the parameter for enhancing the linking essentially depends on the structure of the measuring instrument used and the underlying sample size. Since increasing the density of observed scores with a fixed sample is not possible, one could possibly limit the desired score range when establishing a link function, which would, however, limit the linking only to the central parts of the score distribution. We can observe that small differences between the linking function for males and the linking function for females appeared in the lower (less than -1.8 of the NEPS score values) and upper score range (more than +2.7 of the NEPS score values). Because there were only very few cases in this score range, we decided to consider the full score range.

Regarding the fourth research question, the linking results for the total sample and for the subgroups lead to almost identical distributions of the margins of the PISA proficiency levels. This provides evidence for concordant score distributions of the PISA proficiency levels at the group level.

The classification accuracy at the individual level was only fair. The manifest correlation is r = .68, therefore, the common variance is 46 %. Compared to this value a percentage of 42 % persons who are classified correctly and as far as the low reliable measure allows to see, the classification is not bad. However, it is not the case that the individual students would be classified with a high probability at the same proficiency level with the NEPS score as with the PISA score. Thus, the concordance between both assessments should not be used to calculate the PISA score equivalents of single students.

5.1. Limitations

First, in our study, we used a single group design in which the PISA 2012 mathematics test was administered on the first day and the NEPS 2010 mathematics on the second day. This approach was due to organizational reasons and could not be systematically changed. Therefore, we could not control for an order effect. However, there are two plausible causal interpretations. First, there could be a learning effect that had positive effect on taking the test on the second day. Additionally, a fatigue and demotivation effect could be assumed because students may be more exhausted on the second day compared to the first day. It is possible that both effects cancel one another out (a positive learning effect vs. a negative demotivation effect). However, it is unclear how far this order effect influences the linkage.

Second, we reported the results of an equipercentile equating approach where we used a post smoothing function. We find a high similarity between the smoothed and unsmoothed results, except for some minor differences in the upper and lower score ranges. This result supports the stability of the linking. However, more enhanced linking methods are available. For example, such a further development of the equipercentile linking approach is proposed by Braun and Qian (2007). Their approach contains two modifications, namely, (a) a shift from a school-based to a student-based strategy for estimating the score equivalent to a state standard and (b) the derivation of a more refined estimate of the variance of the score equivalent by considering the

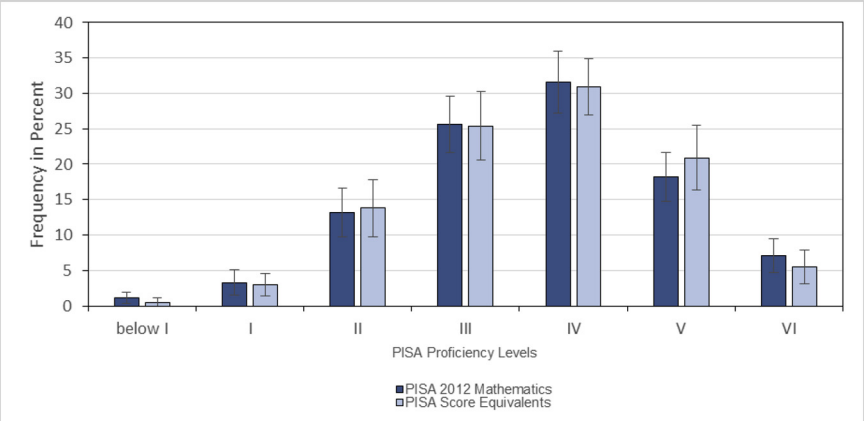

Fig. 6. Classification of male students to PISA 2012 proficiency levels based on the PISA 2012 mathematics scores and the PISA score equivalents.

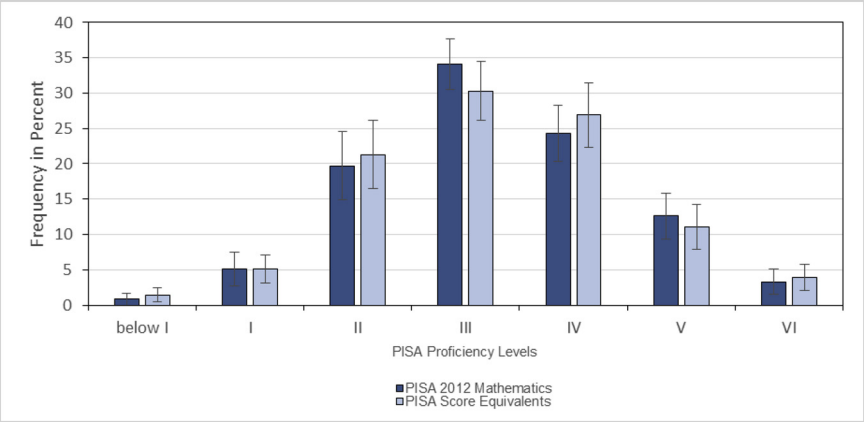

Fig. 7. Classification of female students to PISA 2012 proficiency levels based on the PISA 2012 mathematics scores and the PISA score equivalents.

Table 5
Crosstabulation of students (full sample) to PISA 2012 proficiency levels based on the PISA 2012 mathematics scores and the PISA score equivalents.

		PISA 2012 Mathematics						
		below I	I	II	III	IV	V	VI
PISA Score Equivalents	below I	0.1 %	0.3 %	0.2 %	0.3 %	0.1 %		
	I		1.1 %	2.0 %	0.6 %	0.2 %	0.1 %	
	II	0.5 %	1.9 %	7.2 %	6.1 %	1.6 %	0.3 %	
	III	0.2 %	0.6 %	5.6 %	12.9 %	7.1 %	1.3 %	0.1 %
	IV	0.2 %	0.2 %	1.3 %	8.3 %	12.4 %	5.4 %	1.0 %
	V			0.2 %	1.3 %	5.6 %	6.5 %	2.4 %
	VI				0.2 %	1.0 %	1.7 %	1.7 %

study design (NAEP) in the calculation of sampling error and by obtaining an estimate of the contribution of measurement error (Braun and Qian, 2007, p. 333). Therefore, further research should determine if other linking techniques (also IRT-based methods) would result in similar concordance tables and could therefore validate our findings.

5.2. Conclusions and outlook

Our study provides insight on how national assessments can be related to international assessments, which is an issue that is becoming more important as the participation in both types of assessments has increased in many countries. In the case of our study, the NEPS mathematics assessment has no predefined proficiency levels or other forms of references that yield a criterion-based interpretation of the NEPS mathematics scores. Therefore, using the concordance between NEPS and PISA allows for a mapping of the NEPS mathematics scores to the PISA proficiency levels. These data can be used, for example, for further longitudinal analyses of how the group of underachieving students who belong to the lowest proficiency levels in PISA develops over time. In other cases, when a national and an international measurement scale are linked and both assessments provide their own criterion-based benchmarks, a linking between both reporting scales could be valuable to determine whether, for example, the threshold scores for at-risk students or for the highest benchmark are equivalent or not. This may be the case when local standards within a national assessment are higher than in an international assessment. Cartwright et al. (2003, 2012) provide such examples for Canada. Being able to project scores or proficiency levels from different assessments on one shared scale is also highly useful for benchmarking in educational monitoring (e.g., Kirkpatrick, Turhan, & Lin, 2012). When researchers can put national results from one assessment into the context of an international assessment, policymakers can judge whether an indicator is alarmingly high or not. Furthermore, linking national and international assessments can contribute to prioritizing fields of action regarding instructional and educational practice. For example, when a country finds that the performance of its students is weaker than the student performance seen in neighboring countries and would like to improve the average proficiency, such linking can help countries to take a closer look at any variables worth changing (e.g., grade repetition; see OECD, 2016 and Sälzer, Prenzel, Schiepe-Tiska, & Hammann, 2016).

Acknowledgements

We wish to thank the anonymous reviewers for their large number of hints and suggestions to improve this article.

The study 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.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.stueduc.2020.100847.

References

Adams, R., Wu, M., & Wilson, M. (2012). *ACER ConQuest 3.1 [computer software].* Melbourne: Australian Council for Educational Research.

Blossfeld, H. P., Doll, J., & Schneider, T. (2009). Die Nationale Bildungspanelstudie (NEPS). In W. Böttcher, J. N. Dicke, & H. Ziegler (Eds.). *Evidenzbasierte bildung: Wirkungsevaluation in bildungspolitik und pädagogischer praxis* (pp. 59–68). Münster: Waxmann.

Blossfeld, H. P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: Need, main features, and research potential. *Zeitschrift für Erziehungswissenschaft, 14*, 5–18.

Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.). *Linking and aligning scores and scales* (pp. 313–338). New York: Springer.

Brennan, R. L. (2004). *LEGS: A computer program for linking with the randomly equivalent groups or single-group design.* Iowa City: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Cartwright, F. (2012). *Linking the British Columbia English Examination to the OECD combined reading scale.* Victoria, B.C: Ministry of Education.
Cartwright, F., Lalancette, D., Mussio, J., & Xing, D. (2003). *Linking provincial student assessments with national and international assessments (Education, skills and learning, research papers, No. 005).* Ottawa: Statistics Canada.
Dorans, N. J. (2004). Equating, concordance and expectation. *Applied Psychological Measurement, 28*(4), 227–246.
Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales.* New York: Springer.
Duchhardt, C., & Gerdes, A. (2013). *NEPS technical report for mathematics – Scaling results of starting cohort 4 in ninth grade*NEPS Working Paper No. 22.
Ehmke, T., Köller, O., van den Ham, A.-K., & Nissen, A. (2014). Äquivalenz von Kompetenzmessungen in Schulleistungsstudien. *Unterrichtswissenschaft, 42*(4), 290–300.
Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncomon measures. Equivalence and linkage among educational tests.* Washington, D.C: National Academy Press.
Hambleton, R. K., Sireci, S. G., & Smith, Z. R. (2009). How do other countries measure up to the mathematics achievement levels on the national assessment of educational progress? *Applied Measurement in Education, 22*(4), 376–393.
Hanson, B. A., Harris, D. J., Pommerich, M., Sconing, J. A., & Yi, Q. (2001). *Suggestions for the evaluation and use of concordance results*ACT Research Report No. 2001-1. Iowa City, IA: ACT, Inc.
Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.). *Linking and aligning scores and scales* (pp. 5–29). New York: Springer.
Holmes Finch, W., & French, B. F. (2019). *Educational and psychological measurement.* New York: Routledge.
Huggins, A. C., & Penfield, R. D. (2012). An NCME instructional module on population invariance in linking and equating. *Educational and Psychological Measurement, 31*(1), 27–40.
Kirkpatrick, R., Turhan, A., & Lin, J. (2012). Linking two assessment systems using common-item IRT method and equipercentile linking method. April *Annual Meeting of the National Council of Measurement in Education*.
Kolen, M. J., & Brennan, R. L. (2010). *Test equating, scaling, and linking.* New York, NY: Springer.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310.
Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 4*, 185–207.
Linn, R. L., McLaughlin, D., & Thissen, D. (2009). Utility and validity of NAEP linking efforts. *American Institutes for Research.* URL: https://files.eric.ed.gov/fulltext/ED506806.pdf.
Livingston, S. A. (2004). *Equating test scores. (without IRT) (ETS educational testing service, eds.).* Princeton, NJ.
Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: ETS Policy Information Center.
Muraki, E., Hobo, C. M., & Lee, Y. W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24*, 325–337.
Neidorf, T. S., Binkley, M., Gattis, K., & Nohara, D. (2006). *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Programme for International Student Assessment (PISA) 2003 Assessments (NCES 2006-029).* Washington, DC: U.S. Department of Education. National Center for Education Statistics.
Nissen, A., Ehmke, T., Köller, O., & Duchhardt, C. (2015). Comparing Apples with Oranges? An Approach to link TIMSS and the National Educational Panel Study in Germany via Equipercentile and IRT Methods. *Studies in Educational Evaluation, 47*, 58–67.
OECD (2001). *PISA 2000 knowledge and skills for life: First results from PISA 2000.* Paris: OECD Publishing.
OECD (2014). *PISA 2012 technical report*Paris: OECD Publishing.
OECD (2016). *PISA 2015 results. Policies and practices for successful schools, Vol. II.* Paris: OECD Publishing.
OECD (2010). *PISA 2009 assessment framework - key competencies in reading, mathematics and science.* Paris: OECD Publishing.
OECD (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* Paris: OECD Publishing.
Pommerich, M. (2007). Concordance: the good, the bad, and the ugly. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.). *Linking and aligning scores and scales* (pp. 199–216). New York: Springer.
Pommerich, M., Hanson, B. A., Harris, D. J., & Sconing, J. A. (2004). Issues in conducting linkages between distinct tests. *Applied Psychological Measurement, 28*(4), 247–273.
Sälzer, C., Prenzel, M., Schiepe-Tiska, A., & Hammann, M. (2016). Schulische Rahmenbedingungen der Kompetenzentwicklung. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme, & O. Köller (Hrsg.) (Eds.). *PISA 2015. Eine Studie zwischen Kontinuität und Innovation (S. 176-218).* Münster/New York: Waxmann.
Seidel, T., & Prenzel, M. (2008). Large scale assessment. In J. Hartig, E. Klieme, & D. Leutner (Eds.). *Assessment of competencies in educational contexts. State of the art and future prospects* (pp. 279–304). Göttingen: Hogrefe & Huber.
van den Ham, A. K., Nissen, A., Ehmke, T., Sälzer, C., & Roppelt, A. (2014). Mathematische Kompetenz in PISA, IQB- Ländervergleich und NEPS- Drei Studien, gleiches Konstrukt? *Unterrichtswissenschaft, 42*(4), 321–341.
van den Ham, A. K., Ehmke, T., Nissen, A., & Roppelt, A. (2017). Assessments verbinden, Interpretationen erweitern? Lassen sich die mathematischen Kompetenzskalen im Nationalen Bildungspanel und im IQB-Ländervergleich 2012 verbinden? *Zeitschrift für Erziehungswissenschaft, 20*(1), 89–111.
von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating.* New York: Springer.
Wagner, H., Hahn, H., Schöps, K., Ihme, J. M., & Köller, O. (2018). Are the tests scores of the Programme for International Student Assessment (PISA) and the National Educational Panel Study (NEPS) science tests comparable? An assessment of test equivalence in German Schools. *Studies in Educational Evaluation, 59*, 278–287. https://doi.org/10.1016/j.stueduc.2018.09.002.
Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627.
Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. In H. P. Blossfeld, H. G. Roßbach, & J. von Maurice (Eds.). *Education as a lifelong process. The German national educational panel study (NEPS)* (pp. 67–86). Zeitschrift für Erziehungswissenschaft Special Issue 14.
Wu, M. (2010). *Comparing the similarities and differences of PISA 2003 and TIMSS.* OECD Education Working Papers, No. 32Paris: OECD Publishing.