Majoros *Large-scale Assessments in Education* (2023) 11:14
https://doi.org/10.1186/s40536-023-00162-y

Large-scale Assessments in Education

RESEARCH

Open Access

Linking the first- and second-phase IEA studies on mathematics and science

Erika Majoros[1]*

*Correspondence:
erika.majoros@gu.se

[1] Department of Education and Special Education, University of Gothenburg, P.O. Box 100, 405 30 Gothenburg, Sweden

Reviewing the history of international large-scale assessments (ILSAs), Gustafsson (2008) identified two phases in the work of the International Association for the Evaluation of Educational Achievement (IEA) demarcated by the setup of the new organization in 1990. During the first phase, the IEA conducted separate ILSAs in mathematics and science on four occasions; data were collected on mathematics in 1964 and 1980–82 and on science in 1970–71 and 1983–84. In the second phase, the Third International Mathematics and Science Study in 1995 was the first IEA study to test mathematics and science together. The assessment has been repeated every fourth year, most recently in 2019. Since 1999, the survey is named Trends in International Mathematics and Science Study (TIMSS).

The IEA studies from the first phase have not officially been linked to the TIMSS reporting scale. Previous research has shown that it is possible to link the cognitive outcomes from the two phases of IEA ILSAs on reading and mathematics (Afrassa, 2005; Majoros et al., 2021; Strietholt & Rosén, 2016). However, the studies on mathematics (Afrassa, 2005; Majoros et al., 2021) remained limited in terms of the comparability with the TIMSS reporting scale and the scope of educational systems included in the linking.

Afrassa (2005) used Australian data from the first three IEA ILSAs on mathematics and applied Rasch model equating procedures. Majoros et al. (2021) used data from four countries, England, Israel, Japan, and the USA, and all time points between 1964 and 2015, and applied concurrent calibration using the two-parameter logistic (2PL) model and the generalized partial credit model (GPCM; Muraki, 1992).

The present study aims to link the mathematics and science assessments from the first phase of IEA to the TIMSS reporting scales. This study builds on Majoros et al. (2021) and extends the scope of the linking. Firstly, it uses grade eight data from all participating educational systems. This means 83 educational systems in the studies on mathematics and 85 on science. Secondly, an alternative linking approach is employed to place the results of the first-phase studies on mathematics and science onto the TIMSS trend scale.

The scales achieved by this study may facilitate country-level longitudinal analyses. It is a well-known concern among researchers in the field of social sciences, that due to the cross-sectional survey designs of ILSAs, it is difficult to draw causal inferences about the data (see e.g., Allardt, 1990; Rutkowski & Delandshere, 2016). However, as Gustafsson (2008) pointed out, there are valid approaches to causal interpretation of ILSA data, that have been recently developed, for instance, in the field of econometrics. Suggestions for statistical methods for drawing causal inferences from ILSA data have been made by several researchers (Gustafsson, 2008; Gustafsson & Nilsen, 2022; Robinson, 2013). These powerful analytical approaches such as country-level longitudinal modeling techniques and advanced econometric methods allow for investigating changes in educational systems on the national level or in an international comparative context.

Background

This section is a brief overview of the concept of linking scales, the methodology of scale linking as carried out in TIMSS, and previous research on linking ILSA outcomes.

Linking scales

Kolen and Brennan (2014) defined *equating* as "a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably" (p.2). When statistical adjustments are made to scores on tests that are different in content and/or difficulty, the relationship between scores is referred to as a *linking*, using the terminology of Holland and Dorans (2006), Linn (1993), and Mislevy (1992). The term linking is applied in the present study also following Mazzeo and von Davier (2013), who defined *linking scales* as the process of achieving a scale of results produced by a sequence of assessments, which maintains a stable, comparable meaning over time. In other words, as put by, Dorans et al. (2011) "score equating is a special type of score linking" (p.22).

In the item response theory (IRT) framework, linking multiple tests involves locating the item parameters estimated for the different tests on a common scale. Hambleton et al. (1991) defined four linking designs: single-group designs, equivalent-groups designs, common-persons designs, and anchor test designs. Under the single-group designs, the same group of students completes the tests to be linked. When the tests to be linked are administered to randomly selected groups of students that are equivalent but not identical, the design is referred to as equivalent-groups design. Common-persons designs involve a common group of students across the different groups who complete the tests to be linked. In the anchor test designs, the tests to be linked are presented to different groups of students. Each test includes a set of common items, which set is referred to as the anchor test. This design is commonly used because of its feasibility and avoiding the shortcomings of the single-group, equivalent-group, or common-persons designs, such as fatigue or practice effects.

In the context of IEA ILSAs, the anchor-test design is applied. The achievement tests are completed by different groups of students at different time points and the surveys have maintained a set of common items between consecutive administrations. Within this anchor-test design, there are several possible methods of item calibration for linking the assessments.

The item parameter estimation may be done by pooling the response data from the different tests or in separate calibrations. The first procedure is referred to as concurrent calibration and the item parameters are located on the same scale in one step. This method provides smaller standard errors, involves fewer assumptions than other IRT procedures, and linking may be achieved with few common items (Wingersky & Lord, 1984).

In the case of separate item calibration, one way to link the tests is to calibrate the items on one test and fix the item parameters of the anchor items from this calibration when calibrating the items on the other tests. This procedure is referred to as fixed-parameter calibration (Kolen & Brennan, 2014). Alternatively, the anchor item parameters may be estimated separately followed by a scale transformation so that the distributions of the anchor item parameters match.

Some characteristics of the best equating practices that were described by Dorans et al. (2011) are applicable to scale linking. First, the amount of collected data has a substantial effect on the utility of the resulting equating. Large representative samples ensure that the statistical uncertainty associated with test equating becomes much smaller than other sources of variation in test results. Second, when an anchor test is used, i.e., a set of items that are common between the assessments, the anchor items need to be evaluated via differential item functioning (DIF) procedures to test if they are performing in the same way in the different samples. Finally, the anchor test needs to be highly correlated with the total test score. As Kolen and Brennan (2014) suggested, "the utility and reasonableness of any linking depend upon the degree to which tests share common features" (p.498).

The methodology of scale linking in TIMSS

When it comes to scale linking ILSAs, there are specific additional aspects to be considered. Mazzeo and von Davier (2013) thoroughly described the challenges of designing assessments that can produce comparable results over time and across cultures. One of these challenges is content representativeness, i.e., whether the repeated test material is appropriately representative of the content of the full assessment. The material also needs to be presented and scored in a similar way across assessments. Furthermore, other aspects of the test administration must be sufficiently standardized, i.e., the assessment context needs to be unchanged.

The scale linking of the TIMSS assessments is thoroughly described in the technical reports and what follows is a summary of the procedure (see Martin et al., 2000; Martin et al., 2004; Martin et al., 2016; Martin et al., 2020; Martin & Kelly, 1996; Martin & Mullis, 2012; Olson et al., 2008). The TIMSS scaling procedure involves three steps—the item parameter estimation, the person scoring based on item parameters and population models, and the scale transformation to put generated scores (PVs) on the reporting metric.

The item parameters are estimated via applying concurrent calibration. The concurrent procedure means that the calibration for each TIMSS cycle is based on the pooled data from the current and the previous assessment. A three-parameter logistic (3PL) model is used for binary items, which are multiple-choice items scored as correct or incorrect. A 2PL model is used for binary constructed-response items that have two response options, which also are scored as correct or incorrect. Finally, a GPCM is used for polytomous constructed-response items, i.e., those that are worth more than one score point. Different types of missing data exist in the student test datasets, i.e., omitted, not reached, and not administered. Not-reached items are treated differently in item parameter estimation than in person scoring. In estimating item parameters, these responses are treated as not presented to the students, while they are considered as incorrect responses in person scoring.

The person scoring procedure involves first, drawing five PVs based on the IRT measurement model, and a population model based on contextual data and fit by country. Second, the newly estimated latent ability distribution for the previous assessment data is matched with the distribution of the same data that was estimated in the previous cycle. The final step is to apply this linear transformation to the current assessment data. In the early cycles, the scaling was carried out on equal-sized random samples from each of the participating countries whereas recently, student samples are weighted so that each country contributes equally to the item calibration.

Linking ILSA outcomes

Previous research involving linking cognitive outcomes that are on separate scales in ILSAs over time has applied various linking approaches. Linking can be performed by taking advantage of common items across tests with IRT methods. This approach has been carried out concerning IEA assessments on reading (Strietholt & Rosén, 2016) and mathematics achievement (Afrassa, 2005; Majoros et al., 2021). Furthermore, the IEA recently carried out the Rosetta Stone study, which linked tests of different regional assessment programs for the domains of reading and mathematics with TIMSS and PIRLS. The studies established concordance tables, which project the countries' regional assessment results on the TIMSS and PIRLS reporting scales (Khorramdel et al., 2022a, 2022b). Several other attempts to link test scores from different regional, national, or international assessments over a long period of time have also been made (see e.g., Altinok et al.; Chmielewski, 2019; Hanushek & Woessmann, 2012). These studies rely on IRT within the studies and classical test theory across them because of the limited amount, or lack of, overlapping items.

Against this background, this study was designed to achieve the linking in two main steps. First, the utility of linking the mathematics and science studies administered in the first and the second phase of IEA, that is the current TIMSS scales, is scrutinized by evaluating the degrees of similarity and the behavior of the common items across assessments. The common items that are repeated in succeeding assessments are referred to in this study as *bridge items*. This term was used in the documentation of SISS (Rosier & Keeves, 1991) for common items between the first and second administration of the science ILSAs. The term served as a distinction from *anchor items*, which were used to link the tests across populations. Second, the assessments were placed from the first phase of IEA on the TIMSS reporting scales, including comparing two linking approaches.

Methods

Data

In this study, student data were used from the Third International Mathematics and Science Study (TIMSS 1995) and four ILSAs conducted before 1995: the First International Mathematics Study (FIMS), the Second International Mathematics Study (SIMS), the First International Science Study (FISS), and the Second International Science Study

(SISS). The populations representing 13-year-olds (FIMS and SIMS), 14-year-olds (FISS and SISS), and eighth-grade students (TIMSS 1995) were selected for this study.

All participating educational systems, i.e., countries were included in the analysis. Twelve countries participated in FIMS, 20 in SIMS, 17 in FISS, and 23 in SISS. To improve the comparability of the samples, adjustments were made. For FIMS, this meant keeping students who were in their 7th to 9th year of schooling, and 7th to 10th year of schooling in FISS. Furthermore, cases with missing responses to all items were excluded. After adjusting the samples, in FIMS, 89.95% of the original sample was kept, while in SIMS, 93.62%. In FISS, 87.71% of the original sample was kept, while in SISS, practically 100%. Detailed sample information is available in the documentation of the present study at the COMPEAT repository.[1]

Missing data

Except for FIMS and FISS, in the achievement tests considered in this study, a matrix sampling approach was applied. Matrix sampling of items means that the surveys contain more items in total than what is presented to each student. Consequently, there are item responses in the data that are missing by design. This type of missing data is referred to as *not administered*. However, missing responses may also result from not answering an item, coded as *omitted*. In the TIMSS student achievement tests, if the omitted item is located in a position close to the end of a test booklet, the missing response is classified as *not reached*. Not-administrated items were treated as missing, while omitted responses were treated as incorrect answers both when estimating item parameters and scoring. The not-reached items were treated as missing for item calibration and incorrect responses for student proficiency estimation. This approach is in line with the procedures for handling missing responses in the TIMSS studies. However, in the datasets of the first-phase studies, the various types of missing data were not distinguished and were treated as missing.

Common items

Several common items were identified bridging the first- and second-phase mathematics studies. These items were identical in all tests. The first bridge between FIMS and SIMS consists of 37 items. The second bridge from SIMS to TIMSS 1995 includes 18 overlapping items. In the first bridge, applying the SIMS taxonomy, 15 items were in the arithmetic domain, 10 were in the algebra domain, five were in the geometry domain, two were in the measurement domain, and five were in the statistics content domain. The second bridge covered algebra with three items, arithmetic with six items, geometry with three items, measurement with three items, and statistics with three items.

The number of mathematics items from FIMS to TIMSS 2019, indicating the overlaps is shown in Table 1. Numbers in the same row represent common items between surveys. The number of items common with the preceding administration is shown in the *bridge* row. The total number of items in the item pool per administration is shown in the *total* row. For example, 37 items, which were administered in FIMS, were repeated in

[1] https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat/linking-projects/mathematics-and-science.

Table 1 Number of common and unique items in the respective mathematics assessments

	FIMS	SIMS	T95	T99	T03	T07	T11	T15	T19
	70	37	9	5	3				
		162	9	6	4				
			141	37	16				
				115	56	21			
					115	74	40		
						120	86	52	
							91	76	46
								96	78
									122
Bridge		37	18	48	79	95	126	128	124
Total	70	199	159	163	194	215	217	224	246

This table shows item overlaps across the assessments over time. The TIMSS assessments are denoted by *T* and the last two digits of the year of the assessment cycle. Numbers in the same row represent common items between surveys. The number of items common with the preceding administration is shown in the bridge row. The total number of items in the item pool per administration is shown in the total row. For example, 37 items, which were administered in FIMS, were repeated in SIMS, out of which nine were repeated in TIMSS 1995, then five of these items were also administered in TIMSS 1999, and finally, three of them were included in TIMSS 2003

Table 2 Number of common and unique items in the respective science assessments

	FISS	SISS	T95	T99	T03	T07	T11	T15	T19
	80	32	6	2	1				
		38	7	3	1				
			129	43	22				
				98	52	22			
					123	80	45		
						125	91	52	
							110	94	55
								122	98
									131
Bridge		32	13	48	76	102	136	146	153
Total	80	70	142	146	199	227	246	268	284

This table shows item overlaps across the assessments over time. The same logic applies as in Table 1

SIMS, out of which nine were repeated in TIMSS 1995, then five of these items were also administered in TIMSS 1999, and finally, three of them were included in TIMSS 2003. We can observe that from 1995, no item has been administered on more than three occasions. This is true for the science assessments as well.

Concerning the science studies, in the first bridge from FISS to SISS, 32 identical items were administered as shown in Table 2. Out of these, seven items were classified as life science and 12 as physical Science (Jacobson et al., 1987). According to the technical report, TIMSS 1995 was initially intended to be linked with the results of SISS (Martin & Kelly, 1996). However, formal links between TIMSS and SISS were not established. Investigating the test instruments, 13 items were found to be repeated from SISS to TIMSS 1995 as shown in Table 2. According to the TIMSS 1995 documentation, out of these items, four belonged to life science, two to physics, two to chemistry, two to earth science, and three to environmental issues and the nature of science.

Analytical tools

Data management was done with SPSS 25 (IBM Corp., 2017). IRT analyses were performed with the R package mirt (Chalmers, 2012) for the programming language R (R Core Team, 2022), employing an expectation–maximization algorithm to achieve marginal maximum likelihood estimates of the item parameters outlined by Bock and Aitkin (1981). Latent normal distribution of student proficiency was assumed.

Evaluation of the utility of linking the assessments

The utility of linking the studies was evaluated in two steps. First, the degrees of similarity across assessments were investigated. Second, the behavior of the common items across administrations and in relation to the cross-sectional test was explored. The two first-phase mathematics studies and TIMSS 1995 included sets of common items, i.e., bridges between 1964–1980 (bridge 1), 1964–1995 (bridge 2), and 1980–1995 (bridge 3). Similarly, there were bridges across the science studies between 1970–1984 (bridge 4), 1970–1995 (bridge 5), and 1984–1995 (bridge 6).

The degrees of similarity across the assessments were evaluated based on four criteria suggested by Kolen and Brennan (2014). These criteria are inferences, populations, constructs, and measurement characteristics. Thus first, the measurement goals of the tests to be linked were investigated. Thereafter, the similarity of target populations was considered. Third, the measured constructs were explored in terms of content areas and cognitive domains. Finally, the measurement conditions, such as test length, test format, and administration were evaluated.

The behavior of the common items across administrations was tested first to identify parameter drift using Angoff's delta plot method (Angoff & Ford, 1973) with the deltaPlotR package (Magis & Facon, 2014) for the programming language R (R Core Team, 2022). The choice of method has been made for several reasons. Firstly, the delta plot is a not computationally intensive method. Secondly, this is a relative DIF method, in the sense that items are evaluated with respect to all items. Finally, issues with the traditional DIF analysis, which have been discussed extensively (see e.g., Bechger & Maris, 2015; Cuellar et al., 2021; Doebler, 2019; Yuan et al., 2021), encouraged this choice. Such issues concern the identification problem of IRT parameters (San Martín, 2016) and the circularity problem of obtaining a DIF-free test for estimating the abilities (Cuellar, 2022).

Under this method, the proportion of correct responses is compared between two groups. If there is no item parameter drift, these proportions are located on a diagonal line. Items that are separated from that diagonal are flagged as DIF items. Following the suggestion of Magis and Facon (2014), the threshold was derived by using a normality assumption on the delta points. Each item j has a pair of delta scores (Δ_{j0}, Δ_{j1}), i.e., the delta point. These delta points can be displayed in a scatter plot, referred to as the diagonal plot. The delta scores of the reference group are located on the X-axis and of the focal group on the Y-axis.

The major axis is computed with the following equation:

$$\Delta_{j1} = a + b\Delta_{j0}, \tag{1}$$

in which a is the intercept and b is the slope with

$$b = \frac{s_1^2 - s_0^2 + \sqrt{(s_1^2 - s_0^2)^2 + 4s_{01}^2}}{2s_{01}} \textit{ and } a = \overline{x_1} - b\overline{x_0}, \tag{2}$$

in which $\overline{x_0}$ and $\overline{x_1}$ are the sample means of the delta scores, s_0^2 and s_1^2 are the sample variances, and s_{01} is the sample covariance of the delta scores.

The perpendicular distance D_j between the major axis given in equation (1) and the delta point $(\Delta_{j0}, \Delta_{j1})$, is computed as follows:

$$D_j = \frac{b\Delta_{j0} + a - \Delta_{j1}}{\sqrt{b^2 + 1}}. \tag{3}$$

The other aspect of testing the bridges concerned the relationship of the bridges with the whole test. The correlations of the sum of the correct answers on the bridge items with those on the cross-sectional test for each of the six bridges were tested.

Linking approaches

Two procedures were performed for the mathematics scale. The first approach hereinafter referred to as *four-country-all-time* (points) builds on previous research (Majoros et al., 2021) and uses item parameters calibrated concurrently with data from four countries that participated in every administration up to 2015. The second, hereinafter referred to as the *first-second-time* approach uses item parameters reported for TIMSS 1995 to locate the results of the first and second mathematics surveys on the TIMSS trend scale.

Four-country-all-time

In the four-country-all-time approach, previously (Majoros et al., 2021) estimated item parameters were used, which were calibrated using the pooled data of four countries at each time point from FIMS to TIMSS 2015. The IRT models applied were the 2PL model for dichotomous items, i.e., multiple-choice items and constructed-response items for one score point, and the GPCM for polytomous items, i.e., constructed-response items for two or more score points.

First, the test-takers' abilities were estimated separately for FIMS, SIMS, and TIMSS 1995, by fixing the item parameters to these previously estimated values in the model and drawing five plausible values (PVs) for ability estimates. Then the distribution of the five PVs estimated for TIMSS 1995 was matched with the distribution of the reported TIMSS 1995 PVs. This was done by calculating transformation constants, similarly to the TIMSS scale linking procedure, in two steps.

In the first step, the means and standard deviations of the reported 1995 PVs, which are on the required scale, were matched with the means and standard deviations of the newly estimated PVs for the 1995 data, which are on an independent scale. Then the same transformation constants were employed to rescale the 1980 and the 1964 PVs.

First-second-time

In the first-second-time approach, the item calibration was done by the concurrent calibration of FIMS and SIMS combined with fixed item parameters for the bridge items to TIMSS 1995. The bridge items' parameters were fixed to the values reported for TIMSS

Table 3 Comparison of the amount of data in the linking approaches

	Four-country-all-time					First-second-time
	FIMS	**SIMS**	**T95**	**T99**	**T03**	**T95**
Bridge items	9	17	17	11	7	17
Countries	4	4	4	4	4	42
Sample size	2000	2000	2000	2000	2000	21,000
Responses	18,000	34,000	34,000	22,000	14,000	
Total responses	122,000					357,000

This table shows the number of item responses in the two linking approaches used for the bridge between SIMS and TIMSS 1995. Weighted sample sizes are applied with senate weights that sum up to 500 per country

1995. These item parameters were reported after a rescaling procedure in the 1999 assessment cycle (Martin et al., 2000).

Then the student abilities were estimated separately for FIMS and SIMS, drawing five PVs per test-taker. To locate the student ability estimates on the TIMSS reporting scale, the original transformation constants used for the reported TIMSS 1995 scaling needed to be applied. These constants were acquired through Gonzalez, E. J. (personal communication, September 16, 2022).

For the science scale, the first-second-time approach was chosen for several reasons. First, the IRT models were the same as those used in the TIMSS procedures, i.e., the 2PL and 3PL models and the GPCM. Second, when the amount of information is compared, i.e., the number of item responses used for item calibration in the two approaches, we may note on the one hand that the four-country-all-time concurrent calibration involves 893 items, i.e., all items administered between 1964 and 2015, while the first-second-time approach uses the items administered between 1964 and 1995, i.e., 373 items. On the other hand, the item responses used for the *bridge* between SIMS and TIMSS 1995 are close to threefold in the first-second-time approach (357,000) than those in the four-country-all-time (122,000).

The latter comparison is shown in Table 3, in which weighted sample sizes are applied with senate weights that sum up to 500 per country. The table focuses on the bridge between SIMS and TIMSS 1995 because the approach that uses more information on these items is favorable for better linking. Some bridge items between SIMS-TIMSS 1995 were repeated from FIMS until TIMSS 2003 (see Table 1), hence, in the four-country-all-time approach, responses were used from these surveys.

Weights

In the IRT models, data from different countries contributed equally to the item calibration by applying weights that sum to 500 for each country. In the TIMSS data, this weight variable is referred to as senate weight. In the first-phase studies, weight variables were rescaled to a sum of 500 per country. There were no weight variables in the FIMS datasets; therefore, individuals within a country were weighted equally.

Results and discussion

This section starts with the results of the investigation concerning the utility of the linking. Secondly, the results of the two linking approaches are compared. Then the trend descriptions based on the chosen linking approach are presented. The limitations of the present study are also discussed.

The degrees of similarity across assessments

The following sections are guided by four aspects of the assessments to be linked: inferences, populations, constructs, and measurement characteristics (Kolen & Brennan, 2014). These aspects serve as the criteria for investigating the utility of linking the studies.

Inferences

To evaluate the similarity of inferences drawn from the first- and second-phase IEA assessments, it is useful to distinguish between the levels of inference concerning data, generalization, and explanation (Ercikan & Roth, 2006; Gustafsson, 2008). First, as Gustafsson (2008) pointed out, these assessments use a high-level inference approach to generate data by abstracting information over contexts and items. Second, IEA ILSAs aim to achieve generalizability to the population level by employing sophisticated sampling designs. Finally, these ILSAs were not primarily designed for explanations (Gustafsson, 2018). Overall, the inferences that can be drawn from the IEA studies on mathematics and science are essentially the same in terms of data, generalization, and explanation.

Populations

The populations typically in the 7th–10th year of schooling, i.e., 13- or 14-year-olds were selected in this study. The reason for this selection was that these populations were sampled in all assessments. The target population definitions are shown in Table 4. As can be seen in the table, in the 1980s, the IEA changed the definition of target populations from an age-based to a grade-based for all their studies of student achievement (Strietholt et al., 2013).

In the report on the changes in achievement between the FIMS and SIMS studies, Robitaille and Taylor (1989) argued that the populations targeted across these studies should be considered equivalent. However, between the first-phase science assessments, as Keeves and Schleicher (1992) pointed out, there occurred some sampling deviations. On the one hand, in SISS, there were two options for the target populations. On the other hand, it was decided in most countries that intact classrooms were sampled. To improve comparability, students were selected in this study as closely corresponding to the subsequent studies in terms of grade level as possible as outlined in the Data section.

Constructs

The tests of mathematics and science achievement have been developed based on thorough analyses of the participating countries' national curricula. The items comprising the achievement tests have been selected by specific content areas and cognitive domains. Geometry in mathematics and biology in science are examples of such content areas. The items have been classified to measure different cognitive

Table 4 Target population definitions of the respective studies

FIMS 1964	All pupils who are 13:0–13:11 years old at the date of testing and being in the grade containing the majority of pupils aged 13:0–13:11 years
FISS 1970	All students aged 14:0–14:11 years at the time of testing
SIMS 1980	All students in the grade in which the modal number of students has attained the age of 13:00 to 13:11 years by the middle of the school year
SISS 1984	All students aged 14:0–14:11 years old on the specified date of testing or all students in the grade where most 14-year-old students were to be found on the specified date of testing
TIMSS 1995	All students enrolled in the two adjacent grades that contain the largest proportion of students of age 13 years at the time of testing
TIMSS 1999	TIMSS in 1999 used the same definition as TIMSS 1995 to identify the target grades but assessed students in the upper of the two grades only, the eighth grade in most countries
TIMSS 2003	All students enrolled in the upper of the two adjacent grades that contain the largest proportion of 13-year-olds at the time of testing
TIMSS 2007	All students enrolled in the grade that represents eight years of formal schooling, counting from the first year of ISCED Level 1, provided that the mean age at the time of testing is at least 13.5 years
TIMSS 2011	Same as in 2007
TIMSS 2015	Same as in 2007
TIMSS 2019	Same as in 2007

processes, e.g., knowledge, or reasoning. These processes are referred to as cognitive domains since TIMSS 2003 (Martin et al., 2004).

Since 2007, the TIMSS mathematics and science assessment frameworks have been consistent in terms of content domains. The proportions of items comprising each content domain per administration are shown in Table 7 for mathematics and Table 8 for science in Appendix A. According to the technical report (Olson et al., 2008), mostly organizational revisions were made to the assessment frameworks in 2007. In mathematics, the 2003 *Measurement* domain was eliminated, while the topics covered within that domain were redistributed to geometry or number. In science, the 2003 *Environmental science* domain was eliminated, and its topics were moved to biology and earth science. Overall, the changes in the mathematics and science content domains over time are not substantive but rather terminological. One exception is the science content domain of *Earth science*, which was first introduced in SISS for the 14-year-old population—and then was assessed throughout all time points.

Since 2003, the IEA has been reporting TIMSS achievement results both by content areas and cognitive domains. The cognitive domains “define the sets of behaviors expected of students as they engage with the mathematics and science content” (Martin & Mullis, 2004, p.8). In the administrations preceding 2003, these aspects have been labeled as e.g., the *behavioral categories* (Comber & Keeves, 1973), or *performance expectations* (Martin & Kelly, 1996). These domains were tapping on essentially the same cognitive processes, but the framework has been reconceptualized over time and it has been consistent since 2003. The cognitive domains and the proportions of items per domain are shown in Table 9 for mathematics and Table 10 for science in Appendix A.

Test conditions and instruments

In FIMS (Thorndike, 1967), each student received three booklets with 70 items in total and 60 min time allowance per booklet. Eleven items were constructed-response items, and the rest were multiple-choice items with five response options. Each solved item was worth one score point. In SIMS, (Oldham et al., 1989), matrix item sampling was applied with a core test and four rotated tests. The core test included 40 items to be completed in 35 min. Then the students received one of the rotated tests containing 35 tasks each and were given 40 min to work on it. The overall item pool consisted of 199 multiple-choice items with five response options and each item was worth one score point.

In FISS (Comber & Keeves, 1973), students received two test booklets consisting of 40 items each and 60 min of testing time per booklet. All tasks were multiple-choice items with five response options, for one score point. In SISS (Rosier & Keeves, 1991), matrix item sampling was applied, with one core test and two rotated test booklets, each comprising 10 items. A total of 50 items were presented to each student. The item types and number of response options were the same as in FISS.

In TIMSS 1995 (Martin & Kelly, 1996), students took a test consisting of both mathematics and science items. Matrix item sampling was applied from a pool of 286 (151 mathematics and 135 science) items. Three item types were used, multiple-choice with four or five response options (for one score point), short constructed-response for one score point, and extended constructed-response for two score points In the succeeding cycles of TIMSS, the assessment design has been similar to the one in 1995 (see Martin et al., 2000, 2004, 2016, 2020; Olson et al., 2008; Martin & Mullis, 2012). Some minor changes in the design have occurred over time, for instance, the 2003 assessment was the first TIMSS assessment in which calculators were permitted. In the most recent cycle in 2019, an additional item type was administered, the compound multiple-choice type or multiple selection type.

Parameter drift

The delta plot method was applied for the six bridges across administrations. The plots are shown in Appendix B, in Figs. 4, 5, 6, 7, 8, 9. Two items in the first, one item in the third, and two items in the fourth bridge were flagged for DIF. In the first-second-time approach, these items were treated as unique items in the data instead of bridge items.

Common items and the whole test

The correlations of the sum of the correct answers on the six bridges with those on the whole cross-sectional tests were tested. Preferably, these correlations are high, and the higher the coefficient, the better the anchor test's functioning for linking. The Pearson's correlation coefficients indicate moderate (>0.50) or high (>0.70) positive correlations. In FIMS, (bridge 1) $r = 0.97, p < 0.001$ and (bridge 2) $r = 0.84, p < 0.001$. In SIMS, (bridge 1) $r = 0.88, p < 0.001$ and (bridge 3) $r = 0.66, p < 0.001$. In FISS, (bridge 4) $r = 0.92, p < 0.001$ and (bridge 5) $r = 0.69, p < 0.001$. Finally, in SISS, (bridge 4) $r = 0.86, p < 0.001$ and (bridge 6) $r = 0.80, p < 0.001$.

Table 5 Correlation between FIMS plausible values

First-second-time	Four-country-all-time				
	PV1	**PV2**	**PV3**	**PV4**	**PV5**
PV1	0.910	0.911	0.911	0.911	0.911
PV2	0.909	0.910	0.909	0.911	0.911
PV3	0.910	0.910	0.910	0.910	0.911
PV4	0.910	0.911	0.910	0.911	0.912
PV5	0.909	0.911	0.910	0.911	0.911

All correlations are significant, df = 44,182, p < .001

Table 6 Correlation between SIMS plausible values

First-second-time	Four-country-all-time				
	PV1	**PV2**	**PV3**	**PV4**	**PV5**
PV1	0.917	0.917	0.917	0.917	0.917
PV2	0.917	0.917	0.917	0.917	0.917
PV3	0.917	0.917	0.917	0.917	0.917
PV4	0.917	0.918	0.917	0.916	0.917
PV5	0.917	0.917	0.917	0.916	0.917

All correlations are significant, df = 77,675, p < 0.001

Comparison of the linking approaches

The mathematics plausible scores (five per each test-taker) estimated in the four-country-all-time and the first-second-time approach show strong correlations. The Pearson's correlation coefficients for FIMS are shown in Table 5 and for SIMS in Table 6.

The country means, computed following Rubin's (1987) rules, are shown in Fig. 1 compared by the linking approach. These means consist of all grade levels used in this study because the purpose here is only to compare the results of the two linking methods, not the country results. We may observe that country means are consistently higher in the first-second-time approach with the exception of the low-performing countries in SIMS.

There are three main differences in the linking approaches. First, more item responses were used for the item calibration in the first-second-time approach than in the four-country-all-time approach. This implies more precision of the item parameters. Second, the item calibration is based on data from four educational systems in the four-country-all-time approach, while in the first-second-time approach, data from countries participating in FIMS, SIMS, and TIMSS 1995 were all used, a total of 50 countries. Since in the IRT framework, item statistics are independent of the sample from which they were estimated (Hambleton & Jones, 1993), the differences in the samples should not influence differences in the scores. Finally, in the first-second-time approach, a guessing parameter was included in the IRT model for multiple-choice items. The systematic difference seems to indicate that the IRT modeling mattered in the score estimation differences. The rank order of the countries shows no difference in the two approaches.

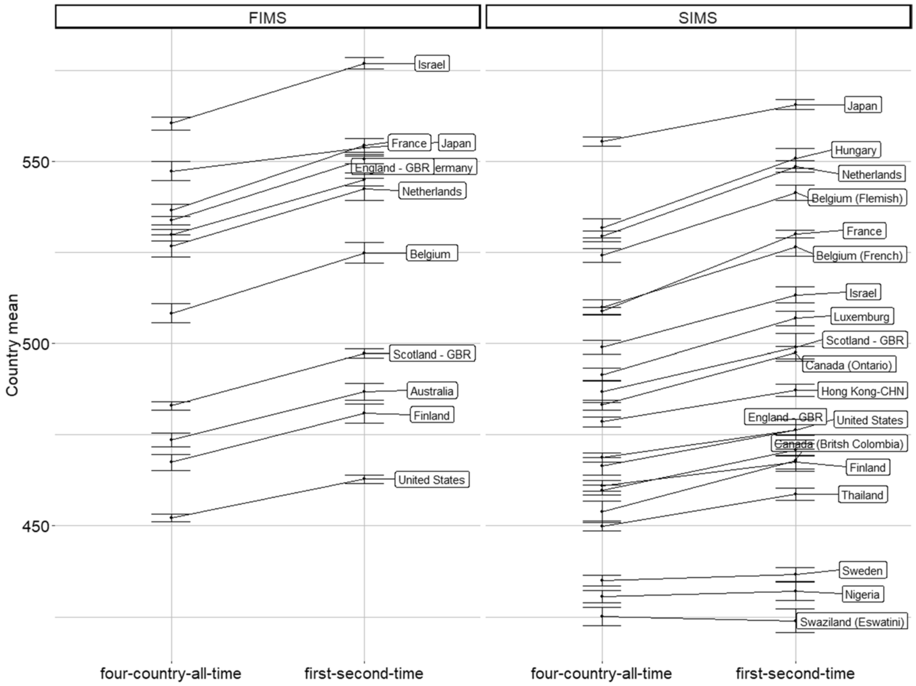

Fig. 1 Comparison of the country means by linking approach

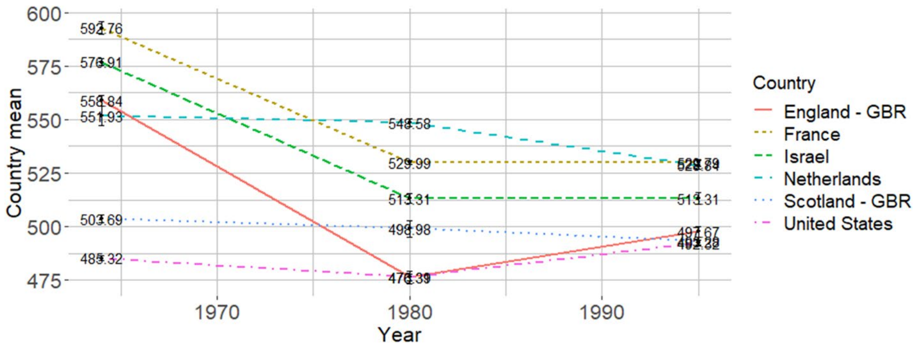

Fig. 2 Trends of grade 8 mathematics achievement

Trend descriptions

Figure 2 shows weighted country means in mathematics for countries that sampled the same grades, i.e., 8 years of schooling in the first-phase studies and TIMSS 1995. This translates to trend descriptions of six educational systems: England, France, Israel, the Netherlands, Scotland, and the United States. Results show a large decline from FIMS to SIMS in the case of three educational systems: France, Israel, and England. The other three systems' performance is rather stable from 1964 to 1980. The country-level changes to 1995 show different patterns in these six countries, a less sharp decline in the Netherlands, no change in France, Israel, and Scotland, and a moderate increase in England and the United States. It may also be seen that the performance of these six countries got closer to each other.

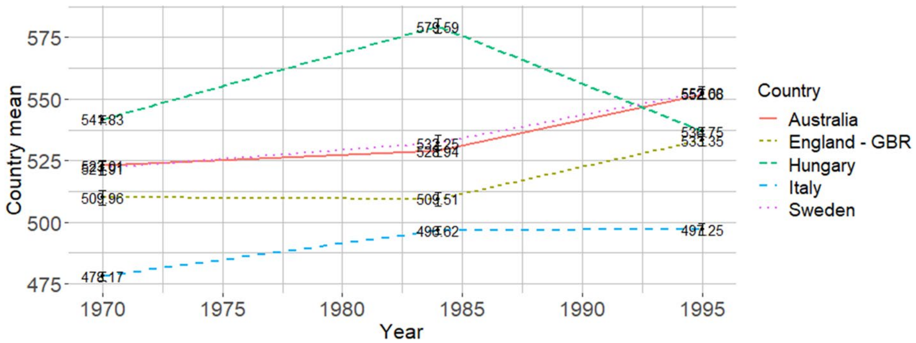

Fig. 3 Trends of grade 8 science achievement

Regarding the science studies, five educational systems sampled the same grades, i.e., eight years of schooling in the first-phase studies and TIMSS 1995: Australia, England, Hungary, Italy, and Sweden. Figure 3 shows the country-level trends in science achievement. All countries showed stable and improving results from the first ILSA to 1995 except for Hungary, which displayed a largely positive, then negative change over time among these countries.

As Mazzeo and von Davier (2013) argued, insufficient content representativeness and/or changes in context can compromise the ability to carry out valid scale linking. Similar to the TIMSS scale linking procedure, all items in each subject domain were calibrated together in this study. Treating the entire mathematics or science item pool as a single domain maximized the amount of data in terms of content representativeness and item responses. The early studies were intentionally designed for measuring change; therefore, this study was carried out under the assumption of sufficient content representativeness across the administrations.

Limitations

The number of common items comprising the bridges from 1980 to 1995 (18 items) and 1984 to 1995 (13 items) is certainly a concern, especially because some of them showed parameter drift and were treated as unique items in the analysis. However, the concurrent calibration method provides the best approach to having only a few bridge items, as pointed out by Wingersky and Lord (1984). They showed that good linking may be achieved with as few as five common items or less with concurrent calibration.

Another limitation concerns the comparability in terms of age and years of schooling. This study used as good approximations of comparable samples over time as possible. In further analyses using the new scale scores, age and grade level can be treated as control variables.

Finally, the coding and treatment of different types of missing data in the achievement tests pose a limitation to this study. In the first-phase studies, the not-reached type of missing responses was not distinguished. Therefore, those missing responses were treated as missing data, unlike in the TIMSS scaling procedure. It would be possible to make this distinction and explore the influence on the results.

Conclusions

In this study, the ILSAs measuring mathematics and science achievement in grade eight from the first phase of IEA were placed on the TIMSS reporting scale. Two linking approaches were compared in terms of the amount of data and the produced scores, extending previous research with more educational systems and subjects from the first phase of IEA ILSAs. The two approaches yielded similar results and the differences might be rooted in the applied IRT models.

The linking was motivated by previous research involving ILSA outcomes that are on separate scales. For instance, Hanushek and Woessmann (2012) used the US National Assessment of Educational Progress to link several ILSAs to the same scale. Their approach assumed that the samples within educational systems are comparable across studies and over time. In contrast, the present study took into account some variations in the comparability of the samples from the participating educational systems over time by applying IRT modeling.

The main purpose was to facilitate future country-level longitudinal studies that include the first-phase IEA studies. Such studies might shed light on explanations for changes in the educational outcomes of participating countries. The results of this study may allow researchers to make reasonable comparisons of these scales over time, even though there have been changes to the instruments, populations, and administration procedures. Making use of modern statistical techniques and reframing the questions and assumptions for the analysis have allowed for the statistical linking of these scales and the possibility of making reasonable comparisons not otherwise available.

Reporting the scales

The first-second-time scales for the first-phase studies are publicly available at the COMPEAT repository[2] along with the documentation of the scale linking. The sampling differences need to be considered when using the scales. For instance, Strietholt et al. (2013) developed a correction model to improve comparability across countries and IEA studies on reading in terms of age and schooling. It is out of the scope of the present study to develop extensive corrections for the sampling composition differences. Another suggestion to account for these differences between time and countries is to treat age and grade level as plausible explanatory variables.

Appendix A

See Tables 7, 8, 9, 10.

[2] https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat/linking-projects/mathematics-and-science.

Table 7 Content areas of the mathematics achievement tests

FIMS 1964	**SIMS 1980**	**TIMSS 1995**
• Arithmetic (47.14) • Algebra (22.86) • Geometry (24.29) • Sets (5.71)	• Arithmetic (31.16) • Algebra (21.11) • Geometry (25.63) Measurement (13.07) • Statistics (9.05)	• Fractions and number sense (33.77) • Algebra (17.88) • Geometry (15.23) • Measurement (11.92) • Proportionality (7.28) • Data representation, analysis, and probability (13.91)
TIMSS 1999	**TIMSS 2003**	**TIMSS 2007**
• Fractions and number sense (37.65) • Algebra (21.60) • Geometry (12.96) • Measurement (14.8) • Data representation, analysis, and probability (12.96)	• Number (29.38) • Algebra (24.23) • Geometry (15.98) • Measurement (15.98) • Data (14.43)	• Number (29.30) • Algebra (29.77) • Geometry (21.86) • Data and chance (19.07)
TIMSS 2011	**TIMSS 2015**	**TIMSS 2019**
• Number (28.11) • Algebra (32.26) • Geometry (19.82) • Data and chance (19.82)	• Number (30.19) • Algebra (29.25) • Geometry (20.28) • Data and Chance (20.28)	• Number (30.33) • Algebra (29.38) • Geometry (20.38) • Data and Chance (19.91)

The percentage of items is shown in parentheses

Table 8 Content areas of the science achievement tests

FISS 1970	**SISS 1984**	**TIMSS 1995**
• Biology (23.75) • Chemistry (23.75) • Physics (27.50) • Practical (25.00)	• Biology (32.86) • Chemistry (21.43) • Physics (32.86) • Earth science (12.86)	• Life science (29.63) • Chemistry (14.07) • Physics (29.63) • Earth science (16.30) • Environmental issues and the nature of science (10.37)
TIMSS 1999	**TIMSS 2003**	**TIMSS 2007**
• Life science (27.40) • Chemistry (13.70) • Physics (26.71) • Earth science (15.07) • Environmental and resource issues (8.90) • Scientific inquiry and the nature of science (8.22)	• Life science (28.57) • Chemistry (16.40) • Physics (24.34) • Earth science (16.40) • Environmental science (14.29)	• Biology (35.51) • Chemistry (19.63) • Physics (25.70) • Earth science (19.16)
TIMSS 2011	**TIMSS 2015**	**TIMSS 2019**
• Biology (36.41) • Chemistry (20.28) • Physics (25.35) • Earth science (17.97)	• Biology (34.09) • Chemistry (20.00) • Physics (25.45) • Earth science (25.45)	• Biology (35.00) • Chemistry (20.00) • Physics (25.00) • Earth science (20.00)

The percentage of items is shown in parentheses

Table 9 Cognitive domains of the mathematics achievement tests

FIMS 1964	**SIMS 1980**	**TIMSS 1995**
• Knowledge and information • Techniques and skills • Translations of data into symbols or schema and vice versa • Comprehension • Inventiveness	• Computation (32.66) • Comprehension (33.67) • Application (28.14) • Analysis (5.53)	• Knowing (21.85) • Performing routine procedures (25.17) • Using complex procedures (21.19) • Solving problems (31.79)
TIMSS 1999	**TIMSS 2003**	**TIMSS 2007**
• Knowing (18.52) • Using routine procedures (23.46) • Using complex procedures (24.07) • Investigating and solving problems (31.48) • Communicating and reasoning (2.47)	• Knowing (33.51) • Applying (47.94) • Reasoning (18.56)	• Knowing (37.67) • Applying (40.93) • Reasoning (21.40)
TIMSS 2011	**TIMSS 2015**	**TIMSS 2019**
• Knowing (36.87) • Applying (39.17) • Reasoning (23.96)	• Knowing (32.55) • Applying (44.81) • Reasoning (22.64)	• Knowing (30.81) • Applying (45.97) • Reasoning (23.22)

The percentage of items is shown in parentheses except for FIMS due to lack of information

Table 10 Cognitive domains of the science achievement tests

FISS 1970	**SISS 1984**	**TIMSS 1995**
• Functional information (21.25) • Comprehension (27.50) • Application (15.00) • Higher processes (11.25) • Practical I (2.50) • Practical III (22.50)	• Information (27.14) • Comprehension (34.29) • Application (38.57)	• Understanding simple information (40.74) • Understanding complex information (28.89) • Theorizing, analyzing, and solving problems (20.74) • Using tools, routine procedures, and science processes (5.93) • Investigating the natural world (3.70)
TIMSS 1999	**TIMSS 2003**	**TIMSS 2007**
• Understanding simple information (39.04) • Understanding complex information (30.82) • Theorizing, analyzing, and solving problems (19.18) • Using tools, routine procedures, and science processes (6.85) • Investigating the natural world (4.11)	• Factual knowledge (30.16) • Conceptual understanding (38.62) • Reasoning and analysis (31.22)	• Knowing (39.25) • Applying (40.19) • Reasoning (20.56)
TIMSS 2011	**TIMSS 2015**	**TIMSS 2019**
• Knowing (33.64) • Applying (42.40) • Reasoning (23.96)	• Knowing (35.00) • Applying (41.36) • Reasoning (23.64)	• Knowing (36.36) • Applying (37.27) • Reasoning (26.36)

The percentage of items is shown in parentheses

Appendix B

See Figs. 4, 5, 6, 7, 8, 9.

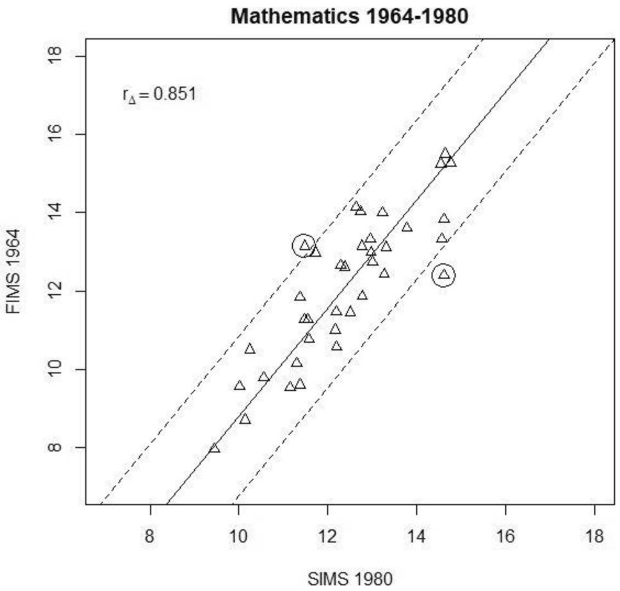

Fig. 4 Delta plot of the bridge between FIMS and SIMS

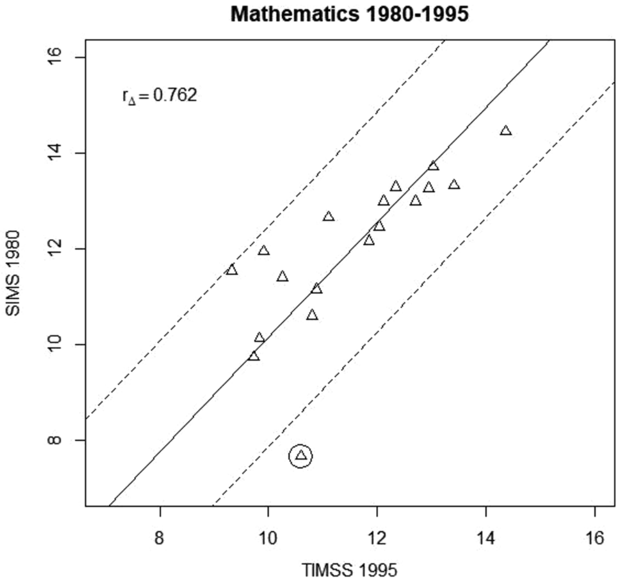

Fig. 5 Delta plot of the bridge between SIMS and TIMSS 1995

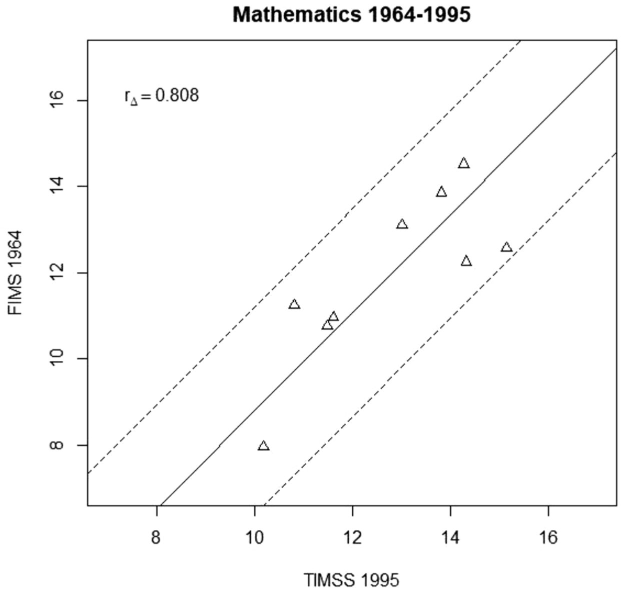

Fig. 6 Delta plot of the bridge between FIMS and TIMSS 1995

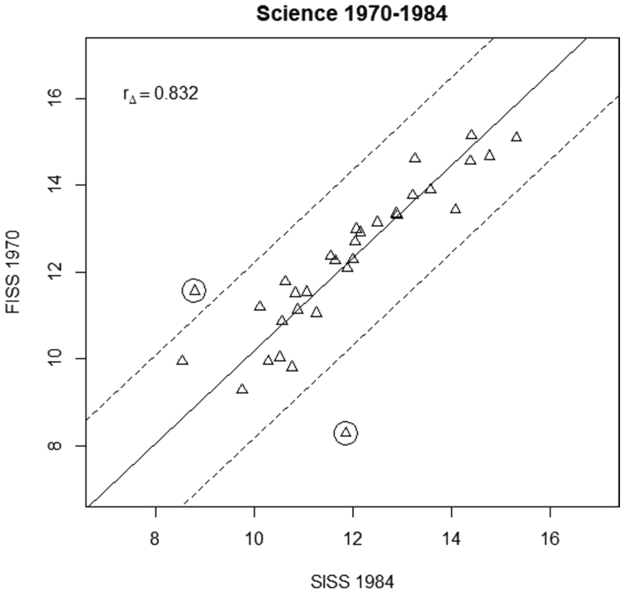

Fig. 7 Delta plot of the bridge between FISS and SISS

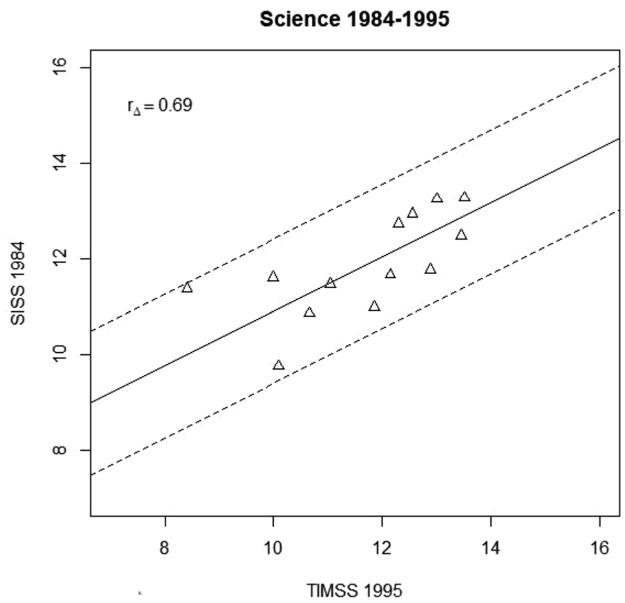

Fig. 8 Delta Plot of the bridge between SISS and TIMSS 1995

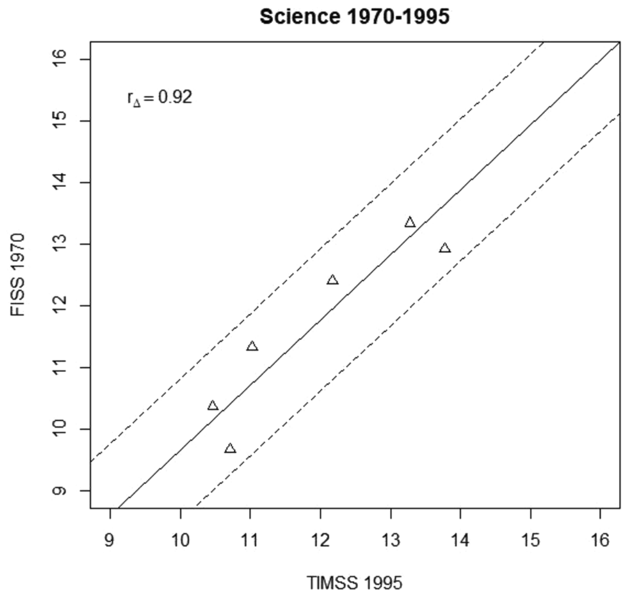

Fig. 9 Delta plot of the bridge between FISS and TIMSS 1995

Abbreviations

2PL	Two-parameter logistic model
3PL	Three-parameter logistic model
DIF	Differential item functioning
FIMS	First International Mathematics Study
FISS	First International Science Study
GPCM	Generalized partial credit model
IEA	International Association for the Evaluation of Educational Achievement
ILSA	International large-scale assessment
IRT	Item response theory

PISA	Programme for International Student Assessment
PV	Plausible value
SIMS	Second International Mathematics Study
SISS	Second International Science Study
TIMSS	Trends in International Mathematics and Science Study

Acknowledgements

In developing the ideas presented here, the author has received great support from Monica Rosén, Jan-Eric Gustafsson, and Stefan Johansson from the University of Gothenburg in Sweden, and Eugenio J. Gonzalez from Educational Testing Service in Princeton, NJ, USA.

Author contributions

The author read and approved the final manuscript.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant agreement no. 765400.

Availability of data and materials

The datasets of the first-phase studies are available in the Center for Comparative Analyses of Educational Achievement repository https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat/studies-before-1995. The dataset of TIMSS 1995 is available in the IEA Study Data Repository https://www.iea.nl/data-tools/repository/timss. The datasets generated during the current study along with the documentation are available in the Center for Comparative Analyses of Educational Achievement repository https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat/linking-projects/mathematics-and-science.

Declarations

Ethics approval and consent to participate

The present study worked with previously collected data of IEA assessments. Therefore, the source data is already anonymized, free, and publicly available. Consequently, ethics approval for this study was not requested.

Consent for publication

Not applicable.

Competing interests

The author reports no competing interests.

Received: 16 March 2022 Accepted: 12 April 2023

Published online: 24 April 2023

References

Afrassa, T. M. (2005). Monitoring mathematics achievement over time: A secondary analysis of FIMS, SIMS and TIMS: A Rasch analysis. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars. Papers in honour of John P. Keeves* (pp. 61–77). Springer.

Allardt, E. (1990). Challenges for comparative social research. *Acta Sociologica, 33*(3), 183–193. https://doi.org/10.1177/000169939003300302

Altinok, N., Angrist, N., & Patrinos, H. *Global data set on education quality (1965–2015): Policy Research working paper; no. WPS 8314*. Washington, D.C. http://documents.worldbank.org/curated/en/706141516721172989/Global-data-set-on-education-quality-1965-2015

Angoff, W., & Ford, S. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*(2), 95–106.

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika, 80*(2), 317–340. https://doi.org/10.1007/s11336-014-9408-y

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459. https://doi.org/10.1007/BF02293801

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v048.i06

Chmielewski, A. K. (2019). The global increase in the socioeconomic achievement gap, 1964 to 2015. *American Sociological Review, 84*(3), 517–544. https://doi.org/10.1177/0003122419847165

Comber, L. C., & Keeves, J. P. (1973). *Science education in nineteen countries: An empirical study. International studies in evaluation: I*. Almqvist & Wiksell.

Cuellar, E. (2022). *Making sense of DIF in international large-scale assessments in education [Doctoral dissertation]*. University of Amsterdam.

Cuellar, E., Partchev, I., Zwitser, R., & Bechger, T. (2021). Making sense out of measurement non-invariance: How to explore differences among educational systems in international large-scale assessments. *Educational Assessment, Evaluation and Accountability, 33*(1), 9–25. https://doi.org/10.1007/s11092-021-09355-x

Doebler, A. (2019). Looking at DIF from a new perspective: A structure-based approach acknowledging inherent indefinability. *Applied Psychological Measurement, 43*(4), 303–321. https://doi.org/10.1177/0146621618795727

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2011). Equating test scores: Toward best practices. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 21–42). Springer.

Ercikan, K., & Roth, W.-M. (2006). What good is polarizing research into qualitative and quantitative? *Educational Researcher, 35*(5), 14–23. https://doi.org/10.3102/0013189X035005014

Gustafsson, J.-E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal, 7*(1), 1–17. https://doi.org/10.2304/eerj.2008.7.1.1

Gustafsson, J.-E. (2018). International large scale assessments: Current status and ways forward. *Scandinavian Journal of Educational Research, 62*(3), 328–332. https://doi.org/10.1080/00313831.2018.1443573

Gustafsson, J.-E., & Nilsen, T. (2022). Methods of causal analysis with ILSA data. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education: Perspectives, methods and findings.* Springer International Publishing.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE Publications.

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth, 17*(4), 267–321. https://doi.org/10.1007/s10887-012-9081-x

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (pp. 187–220). Praeger Publishers.

IBM Corp. (2017). *IBM SPSS Statistics for Windows* (Version 25.0) [Computer software]. IBM Corp. Armonk, NY.

Jacobson, W. J., Doran, R. L., Chang, E. Y. T., Humrich, E., & Keeves, J. P. (1987). *The second IEA science study—U.S.* https://www4.gu.se/compeat/SISS/Design/ED336267.pdf

Keeves, J. P., & Schleicher, A. (1992). Changes in science achievement: 1970–84. In J. P. Keeves (Ed.), *The IEA study of science III: Changes in science education and achievement: 1970 to 1984* (pp. 263–290). Pergamon Press.

Khorramdel, L., Yin, L., Foy, P., Jung, J. Y., Bezirhan, U., & von Davier, M. (2022a). *Rosetta Stone analysis report: Establishing a concordance between ERCE and TIMSS/PIRLS*. TIMSS & PIRLS International Study Center.

Khorramdel, L., Yin, L., Foy, P., Jung, J. Y., Bezirhan, U., & von Davier, M. (2022b). *Rosetta Stone analysis report: Establishing a concordance between PASEC and TIMSS/PIRLS*. TIMSS & PIRLS International Study Center.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*(1), 83–102. https://doi.org/10.1207/s15324818ame0601_5

Magis, D., & Facon, B. (2014). deltaPlotR: An R package for differential item functioning analysis with Angoff's delta plot. *Journal of Statistical Software.* https://doi.org/10.18637/jss.v059.c01

Majoros, E., Rosén, M., Johansson, S., & Gustafsson, J.-E. (2021). Measures of long-term trends in mathematics: Linking large-scale assessments over 50 years. *Educational Assessment, Evaluation and Accountability, 33*(1), 71–103. https://doi.org/10.1007/s11092-021-09353-z

Martin, M. O., Gregory, K. D., & Stemler, S. E. (Eds.). (2000). *TIMSS 1999 technical report.* TIMSS & PIRLS International Study Center.

Martin, M. O., & Kelly, D. L. (Eds.). (1996). *Third international mathematics and science study technical report Design and development.* (Vol. 1). TIMSS and PIRLS International Study Center.

Martin, M. O., & Mullis, I. V. S. (2004). Overview of TIMSS 2003. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 3–21). TIMSS & PIRLS International Study Center.

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. TIMSS & PIRLS International Study Center.

Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 technical report.* TIMSS & PIRLS International Study Center.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS 2015.* TIMSS & PIRLS International Study Center.

Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and procedures: TIMSS 2019 technical report.* TIMSS & PIRLS International Study Center.

Mazzeo, J., & von Davier, M. (2013). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment.* Chapman and Hall/CRC.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. ETS Policy Information Center.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Oldham, E. E., Russel, H. H., Weinzweig, A. I., & Garden, R. A. (1989). The international grid and item pool. In K. J. Travers & I. Westbury (Eds.), *The IEA study of mathematics I: Analysis of mathematics curricula* (pp. 15–53). Pergamon Press.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 technical report.* TIMSS & PIRLS International Study Center.

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria. https://www.R-project.org/

Robinson, J. P. (2013). Causal inference and comparative analysis with large-scale assessment data. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 521–545). CRC Press.

Robitaille, D. F., & Taylor, A. R. (1989). Changes in patterns of achievement between the first and second mathematics studies. In D. F. Robitaille & R. A. Garden (Eds.), *The IEA study of mathematics II: Contexts and outcomes of school mathematics* (pp. 153–177). Pergamon Press.

Rosier, M., & Keeves, J. P. (Eds.). (1991). *The IEA study of science I: Science education and curricula in twenty-three countries.* Pergamon Press.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316696

Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-Scale Assessments in Education*. https://doi.org/10.1186/s40536-016-0019-1

San Martín, E. (2016). Identification of item response theory models. In W. J. van der Linden (Ed.), *Handbook of item response theory: Statistical tools* (Vol. 2, pp. 127–150). Chapman and Hall/CRC.

Strietholt, R., & Rosén, M. (2016). Linking large-scale reading assessments: Measuring international trends over 40 years. *Measurement Interdisciplinary Research and Perspectives, 14*(1), 1–26. https://doi.org/10.1080/15366367.2015.1112711

Strietholt, R., Rosén, M., & Bos, W. (2013). A correction model for differences in the sample compositions: The degree of comparability as a function of age and schooling. *Large-Scale Assessments in Education, 1*(1), 1. https://doi.org/10.1186/2196-0739-1-1

Thorndike, R. L. (1967). The mathematics tests. In T. Husén (Ed.), *International study of achievement in mathematics: A comparison of twelve countries* (pp. 90–108). Almqvist & Wiksell.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8*(3), 347–364. https://doi.org/10.1177/014662168400800312

Yuan, K.-H., Liu, H., & Han, Y. (2021). Differential item functioning analysis without a priori information on anchor items: Qq plots and graphical test. *Psychometrika, 86*(2), 345–377. https://doi.org/10.1007/s11336-021-09746-5

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.