Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 20, Number 14, June 2015 ISSN 1531-7714

Linking Errors Between Two Populations and Tests: A Case Study in International Surveys in Education

Dirk Hastedt & Deana Desa
International Association for the Evaluation of Educational Achievement (IEA)

This simulation study was prompted by the current increased interest in linking national studies to international large-scale assessments (ILSAs) such as IEA's TIMSS, IEA's PIRLS, and OECD's PISA. Linkage in this scenario is achieved by including items from the international assessments in the national assessments on the premise that the average achievement scores from the latter can be linked to the international metric. In addition to raising issues associated with different testing conditions, administrative procedures, and the like, this approach also poses psychometric challenges. This paper endeavors to shed some light on the effects that can be expected, the linkage errors in particular, by countries using this practice. The ILSA selected for this simulation study was IEA TIMSS 2011, and the three countries used as the national assessment cases were Botswana, Honduras, and Tunisia, all of which participated in TIMSS 2011. The items selected as items common to the simulated national tests and the international test came from the Grade 4 TIMSS 2011 mathematics items that IEA released into the public domain after completion of this assessment. The findings of the current study show that linkage errors seemed to achieve acceptable levels if 30 or more items were used for the linkage, although the errors were still significantly higher compared to the TIMSS' cutoffs. Comparison of the estimated country averages based on the simulated national surveys and the averages based on the international TIMSS assessment revealed only one instance across the three countries of the estimates approaching parity. Also, the percentages of students in these countries who actually reached the defined benchmarks on the TIMSS achievement scale differed significantly from the results based on TIMSS and the results for the simulated national assessments. As a conclusion, we advise against using groups of released items from international assessments in national assessments in order to link the results of the former to the latter.

One of the major objectives of international large-scale assessments (ILSAs) is to collect standardized data that allow for cross-national comparisons of student outcomes (achievement, attitudes, etc.) and for examination of the influence of school and classroom factors as well as family background on those outcomes. International assessments conducted on an iterative basis (e.g., on mathematics achievement every three years) enable participating countries to monitor improvement or decline in the achievement of their students. Countries can track trends in that achievement from one assessment cycle to another, from within a grade cohort, and from within national and international contexts. The rich array of information that international assessment programs generates has not only made it possible to describe the educational and social contexts within which students learn but also encouraged many stakeholders to use this internationally comparable information for various purposes including policymaking and decision-making.

National and regional assessments focus on comparing subgroups of students within the country or region. Recently, some countries and regions have attempted to extend these assessments so that the achievement data can be compared with the international achievement benchmarks provided by various ILSAs, such as the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS), both conducted by the International Association for the Evaluation of Educational Achievement (IEA), and the Programme for International Student Achievement (PISA), conducted by the OECD. The procedure is based on the assumption that the items making up the assessment instruments of these large-scale international studies provide a valid link between these studies and the national studies.

In accordance with their policies, IEA and OECD release assessment items and their scoring guides into the public domain after each testing cycle. For instance, IEA releases approximately 40 percent of the TIMSS and PIRLS assessment items, and it keeps the remaining items confidential for the next cycles of the assessment so that they can be used for trends analyses (Mullis, Drucker, Preuschoff, Arora, & Stanco, 2012). IEA releases these assessment materials so that

- Readers of the assessment reports can gain a better understanding of the nature, content, and approach of the assessments; and
- Researchers and other stakeholders can use these materials for research, publication, and teaching purposes (Martin & Mullis, 2012).

The authors of many studies (e.g., Eivers & Clerkin, 2013; Glynn, 2012; Klentschy, 2006; Kosko & Wilkins, 2011; Zonts, 2013) have also found or observed that educators and other stakeholders use the released items to:

- Illustrate how the content of these assessments can be used for educational purposes, for example to conduct a classroom-level investigation into how students' literacy skills impact on mathematics achievement, and what instructional strategies a teacher needs if not all content domains in mathematics have the same degree of relationship with the literacy skills;
- Explore the links between classroom instruction and students' conceptual understandings; and
- Inform discussions about schools' mathematics, sciences, and reading literacy curricula.

The test-development process for each ILSA is guided by an assessment framework that guides multiple reviews and revisions of the assessment items in order to ensure a sufficient number of high-quality items. This rigorous quality-control mechanism is employed because even small flaws in assessment items can accumulate to an extent that they produce skewed or biased findings, a situation not wanted in any international assessment administered to students worldwide and used to inform educational policy decisions.

Today, the practice of including released items from international assessments in national and regional assessments so as to generate average achievement scores that can be linked to the international metric is gaining popularity. However, this practice poses problems, such as those associated with different testing conditions and administrative procedures. It also raises psychometric concerns. For example, the IEA PIRLS and OECD PISA assessment instruments include passages and items embedded in these passages, a design feature that creates the issue of local item dependency and, as a consequence, increased linkage errors (see, in this regard, Monseur & Berezner, 2007; Monseur, Sibberns, & Hastedt, 2008). Earlier studies, such as those by Drasgow (1982), Drasgow, Levine, and Williams (2011), Lamprianou (2010), and Levine and Rubin (1979), point to another problem, evident at the level of the individual test-taker. These researchers claim that test scores that produce unusual response patterns (i.e., spuriously low or spuriously high scores) may not be a valid indicator of an examinee's true ability. As such, a reported test score could be a misleading index.

Our aim in this present study was to shed some light on the effects (linkage errors in particular) we can expect when released items from ILSAs are included in national and regional assessments. We also wanted to investigate differences in linking quality between national/ regional assessments and ILSAs when different sets of test items from the latter are used as the linkage items. While a sound methodology does exist with respect to using test items to link tests, the

process is not always straightforward; its success depends to a large extent on certain factors.

The first relates to how the properties of the linking items function within countries. Although items may have generally satisfactory psychometric characteristics within one country, they may also show differential item functioning (DIF) in other countries. This DIF might be due to content effects (e.g., instructional and curricular variations) and context effects (e.g., wording, position, or exposure) of an item. Second, the linking error[1] may depend on the extent to which the items cover the range of abilities of the examinees in each country. Based on the present simulations, we know that the validity of selecting and using different sets of ILSA-released items in national/regional assessments so that links can be made between the outcomes of both depends on item-by-country interactions, the number of linkage items, and how well the latter match the abilities distribution (i.e., the range of abilities of the population of students taking the test). The linking error and hence the comparability of the achievement scales will also depend on these factors.

Our general objective in the present study was therefore threefold:

1. To gain some idea of the linking quality between ILSAs and national studies of educational achievement by using different sets of TIMSS 2011[2] released test items in simulated national studies;

2. To compare the estimated linking error with that of other ILSAs; and

3. To examine the effect on the extent of linking error when different sets of TIMSS 2011 released items were used as items common to both the national test and the international test.

[1] Linking error can be 'conceptualized as the result of changing the pool of items used to measure achievement as well as shifts in the measurement properties of the common items from one assessment to the next' (Martin, Mullis, Foy, Brossman, & Stanco, 2012, p, 35).

[2] We chose to use TIMSS instead of PIRLS and PISA because of the aforementioned issues of local item dependencies and consequently increased linkage errors in the latter two assessments

Method

TIMSS 2011 assessed student achievement in two subject areas and two grade levels—Grades 4 and 8 mathematics and Grades 4 and 8 science. TIMSS 2011 was the fifth data collection in the TIMSS cycle of studies, and 57 countries participated in it. The tests used to assess student knowledge in the two subjects consisted of both multiple-choice and constructed-response items, the full sets of which were distributed to students according to the TIMSS assessment rotated booklets design (for details, see Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009).

We used the 36 multiple-choice items (see Table 1) that IEA released after completion of the TIMSS 2011 cycle. The Grade 4 mathematics assessment consisted of 180 items in total, 93 of which were multiple-choice.

Table 1. Released items (with DIF and DIF-free) in the test countries: Botswana, Honduras, and Tunisia

Item Identifier		Botswana	Honduras	Tunisia
M031083	M031218	M051091	M031071	M051007
M031071	M031109	M051123	M051007	M041329
M031185	M031159	M041155	M031210	M041155
M051305	M041107	M041320	M031317	M031251
M051091	M041011	M031155	M031155	M031317
M051007	M041041			M031004
M051123	**M041320**			M031043
M051117	M041265			
M041010	M041175			
M041098	M041199			
M041329	**M031210**			
M041158	M031252			
M041155	**M031317**			
M041335	**M031004**			
M041184	**M031043**			
M031187	M031088			
M031251	M031093			
M031294	**M031155**			

Note: Items in bold in the column labeled Item Identifier are those items with DIF in one of the studied countries.

Source: TIMSS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Chestnut Hill, MA.

Taken together, the 180 items assessed such mathematics content domains as number, geometric shapes and measures, and data display. Items also

assessed the cognitive domains of knowing, applying, and reasoning. Documentation about the released items, such as each item's identifier and its content or cognitive domain, along with the relevant scoring guides, can be downloaded at http://timss.bc.edu/timss2011/international-released-items.html.

In order to investigate which conditions produce reliable and valid results when released items from an international assessment are used to link national tests to the international assessment, we estimated the linking error between the international TIMSS 2011 dataset and data from three countries that participated in TIMSS 2011. To do this, we needed to create a national test made up of some of the released items from TIMSS so that the national and international tests shared items in common. We also needed to create the test because at present a national test made up of such items does not exist in reality.

To allow us to link the two tests, we generated 50 sets of binary (multiple-choice) items through a simulation under nine independent conditions that varied according to the number of common items (10, 20, and 30) and according to the presence or absence of DIF in three test countries, Botswana, Honduras, and Tunisia. A description of the procedure we used to select these three countries appears below.

The international group of countries consisted of all countries that participated in TIMSS 2011, minus the country identified as the test country. This meant our study had three international groups, each exclusively defined according to the absence of the country that was the test country. The three groups allowed us to apply a cross-validation process when investigating the extent of linking error from the different populations because the test countries had population characteristics independent of the population characteristics of the three countries on the actual TIMSS 2011 international scale of achievement.

We now describe in greater detail the steps we took to select the test countries and generate data for the simulation study.

Step 1: Selecting the test countries

In order to select these countries, we compared each country that participated in TIMSS 2011 to all other countries that participated in the study. We used the three parameter logistic (3-PL) item response theory (IRT) model in PARSCALE Version 4 (Muraki & Bock, 2003) to conduct the comparisons, which focused specifically on the DIF of the released items in each country. We selected from the countries showing the highest amount of DIF, three countries to serve as the test countries, and assigned a different set of released items to each. The three countries were Tunisia (seven items showed DIF), Botswana (five items) and Honduras (also five items). We chose these countries because all are developing countries and there presently seems to be an increasing interest in such countries of linking national assessments to international assessments. Table 2 lists, for each test country, the released items that showed DIF and those that did not.[3]

Table 2. Population and common item characteristics used for the simulation of the national tests

Group (*N*)	Population	10*	20*	30*
Tunisia (5,000)	N(-.726,.884)	7 DIF, 3 DIF-free	7 DIF, 13 DIF-free	7 DIF, 23 DIF-free
Botswana (4,000)	N(-1.535,.947)	5 DIF, 5 DIF-free	5 DIF, 15 DIF-free	5 DIF, 25 DIF-free
Honduras (4,000)	N(-.963,.882)	5 DIF, 5 DIF-free	5 DIF, 15 DIF-free	5 DIF, 25 DIF-free

Note: *Common items are the released items from TIMSS 2011. There were sets of 60, 50, and 40 unique items in the national tests, and the total number of test items is 70.

In Botswana and Honduras, four of the 36 TIMSS 2011 Grade 4 mathematics multiple-choice items exhibited DIF, but these four items were different in each country. However the two countries also had one other released item with DIF, and this item—Item M031155, see Table 1—was the same in each country. Item M041155 appeared to show DIF in both Botswana and Tunisia, and Item M051007 showed the presence of DIF in Honduras and Tunisia. None of the listed items that showed DIF appeared in all test countries.

[3] Summary of the PARSCALE DIF test results for each country is available upon request.

Step 2: Defining the test forms and data generation

Our next step was to use the item and population (i.e., the students who took the test) characteristics of the test countries to generate simulated national tests, that is, one national test for Botswana, one for Honduras, and one for Tunisia. The characteristic that each national and international TIMSS test had in common was the set of released items.

In order to investigate linking error, we used three such sets of released items for each country. The sets varied in terms of the number of items held in common. Set 1 contained 10 of these items, Set 2 contained 20, and Set 3 contained 30. We used all released items showing DIF in each test country as common items because the number of released items with DIF would always be lower than the number of common items set for the national simulation tests (see Table 2). For example, for Tunisia, although the numbers of common items examined were 10, 20, and 30, each of these sets contained seven items with DIF.

We used the released items not only to link the two tests but also to estimate the IRT proficiency scores for the national and international populations. The unique items were the items that were administered exclusively to the national test or to the international TIMSS 2011 but not both. In order to ensure that the simulated item parameters used to generate the response data resembled the parameters likely to be found in TIMSS 2011, we endeavored, to the greatest extent possible, to keep the population characteristics of the test countries (i.e., number of examinees and ability distribution) and their associated released item characteristics (item difficulty, discrimination, and lower asymptote parameters) the same as they were in the international assessment. To do this, we used the 3-PL IRT model to generate binary items. An R program generated the data, wrote the command files for the output from PARSCALE, and also executed and processed that output. A summary of the population characteristics of the test countries can be found in Table 2.

The total number of items in each simulated national test was 70; the observed international test had a total of 93 items. All items were multiple-choice and were scored 0 for an incorrect or 1 for a correct answer. We carried out 50 replications per each simulation set per country, which produced 450 datasets in total (i.e., 50 replications x 3 sets of tests x 3 test countries).

Step 3: Estimating the linking error

The linking procedure that we used (within the IRT framework) was the non-equivalent groups anchor test (NEAT) (Kolen & Brennan, 1987, 2014). As mentioned above, we used the 3-PL IRT model to estimate the population and test item parameters. We then carried out IRT calibrations of the international and national tests. The calibrations for the simulated national tests data were replicated 50 times per set of released items per test country. The procedures involved in linking the two tests consisted of three stages:

1. Using the separate calibrations to carry out IRT calibrations of the national tests and TIMSS 2011 international achievement scale;
2. Estimating proficiency scores for each national test on the basis of the TIMSS 2011 international achievement scale; and
3. Placing the predicted proficiency (IRT) scores from the national tests on the estimated TIMSS 2011 results.

More specifically, after having conducted each separate calibration of the national and international tests (Step 1), we used the Stocking-Lord characteristics curve method (Kim & Lee, 2004; Kolen & Brennan, 2014; Ogasawara, 2001) to transform and compute the linking constants (Step 2). We then used the NEAT procedure to transform the estimates from the national tests onto the international TIMSS 2011 scale (Step 3), so that the scores from the national tests were in the same metric as those in the international test.

For each pair of linking results (i.e., from the national test to the international scale), we computed the linking error from the standard deviation of the national test proficiency scores obtained after placement on the international TIMSS 2011 scale. The linking error was therefore computed as

$$LE = \sqrt{\frac{var(\theta_{shift})}{n}} \qquad (1)$$

where $var(\theta_{shift})$ is the variability of the differences in test scores and n is the number of (common) released items.[4]

Results and discussion

Our analysis produced three major findings, brief descriptions of which follow. We also consider the implications of these and several other findings for countries wanting to use items from ILSAs as linking items in their national assessments.

Main findings

1. *The linking errors for each test country were substantially larger than the average linking errors for other selected ILSAs:* Table 3 presents the results of the linking error computations for each test country, while Figure 1 plots the estimated linking errors for each country. Figure 1 also shows how the plotted linking errors compared with the international average linking errors reported for three other ILSAs—PIRLS, PISA, and the TIMSS/National Assessment of Educational Progress (NAEP) survey.[5] Here we can see that the average linking errors from our study were noticeably high in eight out of the nine cases. They were all above the average with respect to the PISA mathematics test, above the average for the PIRLS reading assessment, and above the average for the TIMSS/NAEP mathematics test.

Table 3. Linking error per set of released items per test country

Number of released items	Test countries		
	Botswana	Honduras	Tunisia
10	9.12	10.11	6.47
20	2.78	5.30	5.03
30	1.07	2.84	6.07

2. *The number of common items in the different sets of released items appeared to affect the size of the linking error*: We can also see from Figure 1 that all three of the current study's linking errors were larger than the average ILSA cutoffs and that the highest observed linking errors were those for the set of 10 common released items (LE of 9.1 for Botswana, 10.1 for Honduras, and 6.5 for Tunisia). Furthermore, even though the size of the linking errors for the three countries decreased as the number of common released items increased to 20, they were still larger than the ILSA averages. By the time the number of common released items reached 30 items, the linking error for Botswana (LE of 1.1) had dropped to be comparable with the ILSA averages but the linking errors for the two other cases were still larger (Tunisia with an LE of 6.0, and Honduras with an LE of 2.8).

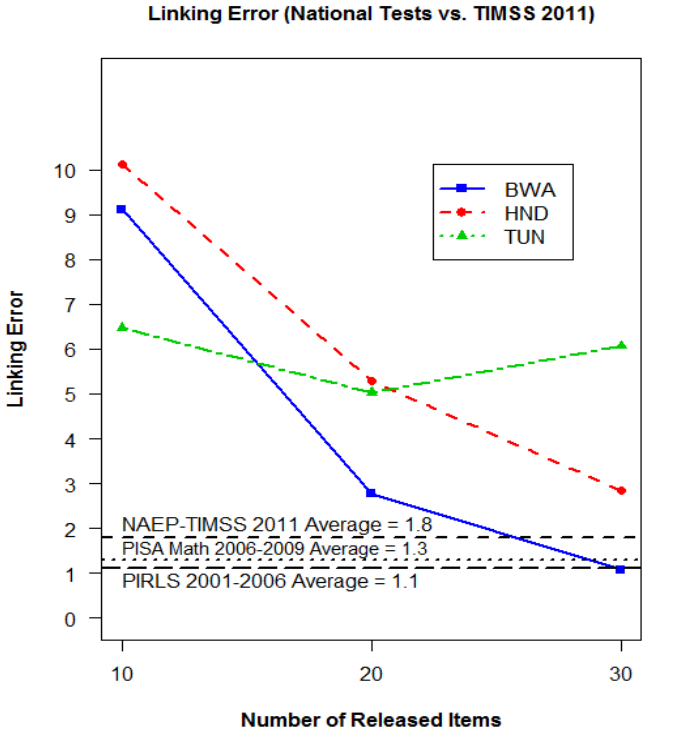

Figure 1. Linking errors for the test countries and the international averages in PISA (Grade 4 mathematics), PIRLS (reading literacy), and NAEP/TIMSS (Grade 8 mathematics)

3. *Populations of students with lower average proficiency scores and higher numbers of released DIF items produced a stable larger amount of linking error, regardless of the number and selection of items common to both the national and international tests*: The average TIMSS 2011 mathematics achievement scores of the populations of Grade 4 students in the test countries were all lower than the TIMSS 2011 international average proficiency score. As noted earlier, these three countries were all countries where students' mean proficiency during the actual

[4] The formula is adopted from Martin et al. (2012), and Monseur and Berezner (2007).

[5] Details can be obtained from Jia et al. (2014), Martin, Mullis, Foy et al. (2012), and OECD (2012).

TIMSS 2011 assessment placed them at the lower levels of the international achievement scale. They were also countries where, according to our computations, the number of released items exhibiting DIF was relatively high. For example, in Tunisia, where the mean of the ability estimates (in logits) was -.726 and seven released items showed the presence of DIF, the linking errors were larger than the averages of the ILSAs depicted in Figure 1 regardless of which set of released items was used in the national test.

Linking errors and descriptive results

In general, we can assume that there is good linkage between international and national test items when there is no significant difference between the average achievement scores computed from the two test administrations. However, where this is not the case, there is the danger of over-estimating examinees' scores. Consider, for example, the test in this study with 10 common items. Here, every examinee's score would be overestimated by ten, nine, or six score points for Honduras, Botswana, and Tunisia, respectively, on the TIMSS 2011 international achievement scale for Grade 4 mathematics.

Table 4 presents the mean TIMSS 2011 achievement scores reported for the test countries. It also shows the means for the national tests in the current study for each of the three sets of released items, and the deviations of the national means from the international means. As an example, let us consider the comparisons for Case 1, the 10 released items. During the TIMSS 2011 assessment, students in Botswana, Honduras, and Tunisia achieved a proficiency score on the international scale of 419 (SE =3.7), 396 (SE =5.5), and 359 (SE =3.9), respectively. In the national tests, the respective scores were 442 (SE =1.4), 434 (SE =1.3), and 449 (SE =1.3). The mean scores of the national tests were thus overestimated when compared to the 2011 TIMSS scaling, by 23, 38, and 90 score points for each of the test countries.

The mean differences between the international and national tests and their standard errors (i.e., linking errors) in the last column of Table 4 are the sums of the squared deviations of scores obtained when the national tests were scaled on the TIMSS 2011 international achievement scale. By taking linking errors into account, we can compute the standard errors of these differences as

$$SE = \sqrt{var_{TIMSS} + var_{National} + var_{linking}} \quad (2a)$$

Hence, the standard errors of the differences are

$$SE_{BWA} = \sqrt{3.7^2 + 1.4^2 + 9.1^2} = 9.923 \quad (2b)$$

$$SE_{HND} = \sqrt{5.5^2 + 1.3^2 + 10.1^2} = 11.574 \quad (2c)$$

$$SE_{TUN} = \sqrt{3.9^2 + 1.3^2 + 6.5^2} = 7.691 \quad (2d)$$

The standardized difference, t, for each test country is therefore as follows

$$t_{BWA} = 23/9.923 = 2.318 \quad (3a)$$

$$t_{HND} = 37/11.574 = 3.242 \quad (3b)$$

$$t_{TUN} = 90/7.691 = 11.702 \quad (3c)$$

and each is statistically significant (values > 1.96) at the 95% confidence level. Across the three cases in Table 4, only one difference (in Case 2 for Botswana) was not significant, indicating equivalency between the national test score and the TIMSS 2011 scaling.

Table 4. Means (standard errors of the mean) and deviations (linking error) for TIMSS 2011 and the national tests

Test country	Estimated from TIMSS 2011(s.e)[a]	Estimated from national tests (s.e)[b]			Means difference (LE)[c]		
		Case I (10 items)	Case 2 (20 items)	Case 3 (30 items)	Case 1 (10 items)	Case 2 (20 items)	Case 3 (30 items)
Botswana	419 (3.7)	442(1.4)	412(1.3)	410(1.3)	23(9.1)*	7(2.8)	9(1.1)*
Honduras	396 (5.5)	434(1.4)	374(1.2)	376(1.2)	38(10.1)*	22(5.3)*	20(2.8)*
Tunisia	359 (3.9)	449(1.3)	380(1.1)	409(1.2)	90(6.5)*	21(5.0)*	50(6.1)*

Notes:
[a] National test on the 2011 TIMSS scale (From Exhibit 1.1 in T11_IR_Mathematics_FullBook.pdf)
[b] The national test scores based on the national test item parameters
[c] Deviation of scores when the national test of the test country was scaled on the 2011 TIMSS (i.e., errors); *t>1.96.

From Table 5, we can see that all pairwise comparisons among the countries were inconsistent with the significance difference computed from TIMSS 2011. The 10-common-item design failed to detect the differences among the countries, and the 20- and 30-common-item designs showed varying results, thus implying that only the TIMSS scores correctly detected the differences in achievement across the three countries. Because all of the scores, except one, derived from the national tests deviated significantly from the TIMSS 2011 results, we can anticipate that this would affect the statistical significance of the trend estimate, namely, that the results of the benchmark levels would remain significantly different as described below.

Table 5. Mean, standard error, and differences in the means of the two tests

Test	Botswana	Honduras	Tunisia	Botswana-Honduras[1]	Botswana-Tunisia[1]	Tunisia-Honduras[1]
TIMSS	419 (3.7)	396 (5.5)	359 (3.9)	23(5.4)*	60(6.6)*	37(6.7)*
Case 1 (10 items)	442(1.4)	434(1.4)	359 (3.9)	9(13.8)	6(11.3)	15(12.2)
Case 2 (20 items)	412(1.3)	374(1.2)	449(1.3)	38(6.2)*	32(6)*	6(7.5)
Case 3 (30 Items)	410(1.3)	376(1.2)	380(1.1)	34(3.5)*	1(6.4)	33(6.9)*

Notes:
* Significant at .05 level.
[1] The standard error for the country-TIMSS difference was calculated using the regular computation, and with the standard error for the differences between countries taking into account the linking error provided in Table 4.

Linking errors and benchmark trends

TIMSS 2011 established four international benchmarks on its international achievement scale for Grade 4 mathematics. These were advanced (scale score 625), high (550), intermediate (475), and low (400). When we examined student achievement at the low international benchmark, we found all results for the national tests differed significantly from those for the TIMSS assessment. This pattern is evident in Figure 2, where none of the 95% confidence intervals (error bars of the percentage) for the 10-, 20-, and 30-item national test designs overlap with the interval for the TIMSS 2011 test.

Table 6 presents the percentages of students in each country and for each item set that reached the assessment benchmarks on the TIMSS 2011 Grade 4 mathematics achievement scale. It also allows us to see whether these percentages aligned with the benchmarks the students (in percentages) from the three countries reached in reality when they took the TIMSS international test (see in this regard Exhibit 2.2 of Chapter 2 in Martin & Mullis, 2012). In all cases, the percentages for the national tests differed significantly from the TIMSS' percentages.

We suggest that any country developing national tests from items in international assessments should take into account these findings and their implications. First, for all cases, the national benchmarking at the international levels followed the same direction in terms of more examinees reaching the lower than the higher benchmarks. Second, the national tests overestimated the international benchmark at the advanced level in all cases (except one case for Honduras), and overestimated all benchmarks for the case in which the national tests had 10 released items in common with the international assessment. We consider these two findings have the following implications:

1. All of the test countries (or parents of the students) can claim (or think) that their students are actually doing well because some of the students performed proficiently in the national test despite the whole cohort of students having performed poorly when compared internationally. This situation could lead to a country's government officials, policymakers, and members of the public reaching misleading conclusions about their students' actual levels of proficiency.

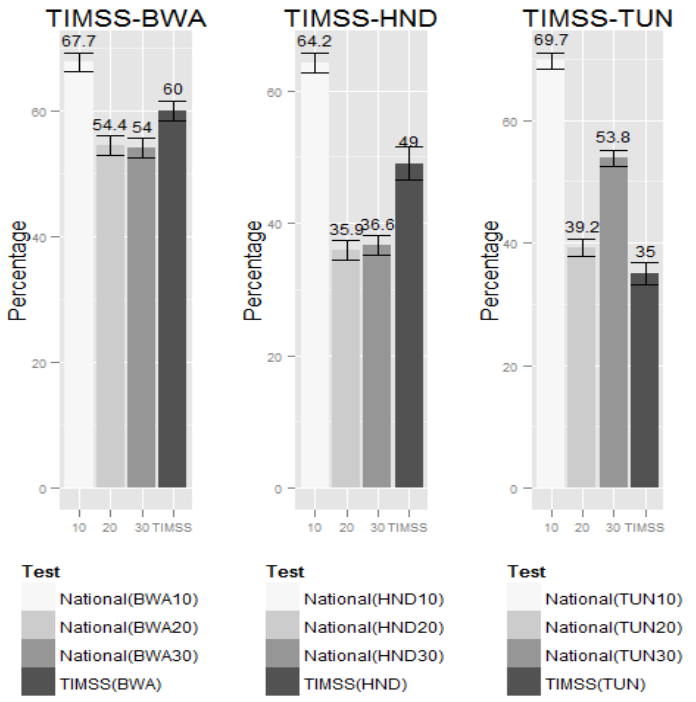

Figure 2. Low international benchmark (percentage, standard error) for all cases per country

2. Reporting the results of national tests may accordingly be favored over reporting the results of the international tests, but this practice could deny students opportunity (e.g., remedial education programs) to attain the proficiency levels they need to study successfully at higher education levels.

A third consideration aligns with our finding that the Case 2 and Case 3 national tests (20 and 30 released items in common) in Botswana and Honduras underestimated the percentages of students reaching the high, intermediate, and low benchmarks. This finding also has two implications.

1. The proficiency benchmarks that students reach when tested nationally may be higher than the benchmarks they attain when tested within a global context.

2. The more similar the national tests are to the international assessment, the more likely it is that student proficiency will be observed as low at the national level. This means that the linking quality of the tests may still not be sufficiently robust to allow meaningful comparisons between performance on the national test and performance on the international assessment.

Our fourth point concerns our finding that the national tests in Honduras, where the population and item characteristics were below the TIMSS 2011 data averages and where there were more released items with DIF than in the other two countries, showed unstable benchmark levels. At the advanced and low levels, the percentages of Honduran students achieving the benchmarks were all overestimated from the international trends. At the middle levels (i.e., high and intermediate), four out of six cases from the national tests with 10 or 30 released items overestimated the international benchmarks, while the national test with 20 released items underestimated the international benchmarks. This finding has implications similar to those already mentioned, with the performance of students potentially becoming less tractable as the benchmark levels become more contradictory. The different benchmark levels accordingly make it difficult to conduct an in-depth evaluation as to what level of proficiency students in each country are actually reaching in mathematics internationally.

Table 6. Percentages (standard errors) of students reaching the international benchmarks

	Low	Intermediate	High	Advanced
BWA (TIMSS)[1]	60(1.6)	29(1.7)	7(1.1)	0(0.1)
BWA (10 Items)	68(1.5)	35(1.5)	11(1.0)	2(0.4)
BWA (20 Items)	55(1.5)	22(1.3)	4(0.6)	1(0.2)
BWA (30 Items)	54(1.5)	21(1.3)	4(0.6)	1(0.2)
HND (TIMSS)[1]	49(2.5)	29(2.1)	7(0.8)	0(0.1)
HND (10 Items)	64(1.5)	32(1.4)	9(0.9)	2(0.4)
HND 20 Items	36(1.5)	9(0.9)	1(0.3)	0(0.0)
HND (30 Items)	37(1.5)	10(0.9)	1(0.3)	0(0.1)
TUN (TIMSS)[1]	35(1.8)	11(1)	2(0.3)	0(0.0)
TUN (10 Items)	70(1.3)	38(1.4)	13(0.9)	3(0.4)
TUN (20 Items)	39(1.4)	10(0.8)	1(0.3)	0(0.1)
TUN (30 Items)	54(1.4)	20(1.1)	4(0.5)	0(0.2)

Note: [1] National test on the 2011 TIMSS assessment (from Exhibit 2.2 in T11_IR_Mathematics_FullBook.pdf); All national-TIMSS percentage paired comparisons are significantly different at .05 level; percentage is rounded to the nearest whole number.

Summary and conclusions

Our study showed that, in two out of the three national cases, the linkage error decreased significantly as the number of common items in the national tests increased. For Tunisia, the linkage error remained at a high level, but for Botswana the linkage error decreased from nearly nine score points for the 10-item case to a level comparable to the linkage error in the international assessment for the 30-item case. For Honduras, the linkage error decreased from more than 10 score points for the 10-item case to an amount about twice the size of the linkage error found in the international assessment. However, in nearly all cases, the differences across the three countries would be statistically significant only with respect to the TIMSS international results. The results based on the national assessments would not be able to detect these differences.

When comparing the country averages calculated on the simulated national assessments with the international TIMSS results, we found that the results for Botswana came within the range of 10 score points for the cases with 30 and 20 common items. Interestingly, the results were no better with 30 than with 20 items. For Honduras, the difference between the simulated national assessments and the international TIMSS assessment was 38 score points in the case of 10 common items, with that difference decreasing to about 20 score points in the 30- and 20-item cases. The 10- and 30-item cases for Tunisia showed the greatest differences from the international averages, with 90 and 50 score points, respectively; the difference in the average for the 20-item case was smaller, at 21 score points.

When we compared the estimated country averages based on each simulated national test and the average based on the international TIMSS assessment, we found that the estimates came relatively close to each other in only one instance across the three countries. We furthermore found that the estimates for the percentages of students reaching the defined benchmarks on the TIMSS scale differed significantly from the benchmarks for TIMSS and the benchmarks for the simulated national assessments.

The results presented in this study are consistent with the literature showing that sampling common items can produce a substantial source of error for the ability estimates (see, for example, Haberman, Lee, & Qian, 2009; Michaelides & Haertel, 2004; Monseur & Berezner, 2007; Xu & Davier, 2010). Because obtaining valid test scores is vital for valid interpretations or comparisons of assessment outcomes, the sets of common items linking two assessments should have minimal effects on response outcomes.

However, in conclusion we do not feel it is wise to recommend using groups of released items from international assessments in national assessments in order to provide a link to the results from the latter. While it is possible for some results of national tests to nearly approximate the international results, as happened for Botswana in this study, there is also the likelihood of the results differing significantly as occurred for Honduras and Tunisia. However, such differences might not even be detected, as occurred with Honduras. There, the standard errors decreased significantly as the number of linkage items increased, but the estimated results from the national survey still fell far short of the international results.

Limitations

The simulation in the present study included three of the lower-achieving countries that participated in the TIMSS 2011 Grade 4 mathematics assessment. If we had looked at higher achieving countries, the results and the conclusions drawn might differ from what we have presented here. While this matter merits further evaluation, current discussions about the need to raise student achievement in lower-achieving developing countries, many of which do not participate in international assessments, warrants keeping the focus on such countries. It is possible that linkage problems in countries achieving at even lower levels than the three considered in this paper would generate even more problematic findings, and so this consideration also warrants investigation.

In addition, common-item sampling replication methods (e.g., Jackknife and bootstrap) may offer an interesting alternative to empirically DIF-oriented simulation because these methods would probably be better suited to the complex structure of TIMSS (e.g., the balanced incomplete block design and stratified student sampling). Future research could also take into account the inclusion of examinee-sampling error that can arise from one or both of two factors: the sampling of examinees and the sampling of items (Haberman et al., 2009; Johnson, 1989; Sheehan & Mislevy, 1988).

While the results of the present study usefully indicate the need to address the number of released items when linking two tests with the aim of enabling comparison between national and international trends data, using sets of released items from the international assessment to link the two tests may lead to misspecification and misinterpretation of student achievement at the national level.

References

Drasgow, F. (1982). Appropriateness measurement. *Applied Psychological Measurement*, 6(3), 297–308.

Drasgow, F., Levine, M. V., & Williams, E. A. (2011). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1), 67–86.

Eivers, E., & Clerkin, A. (2013). National schools, international contexts: Beyond the PIRLS and TIMSS test results. Dublin, Ireland: Educational Research Centre.

Glynn, S. M. (2012). International assessment: A Rasch model and teachers' evaluation of TIMSS science achievement items. *Journal of Research in Science Teaching*, 49(10), 1321–1344. http://doi.org/10.1002/tea.210595

Haberman, S., Lee, Y.-H., & Qian, J. (2009). Jackknifing techniques for evaluation of equating accuracy (No. ETS RR-09-39). Princeton, NJ: Educational Testing Service (ETS).

Jia, Y., Phillips, G., Wise, L. L., Rahman, T., Xu, X., Wiley, C., & Diaz, T. E. (2014). 2011 NAEP-TIMSS linking study: Technical report on the linking methodologies and their evaluations (No. NCES 2014-461). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Johnson, E. G. (1989). Considerations and techniques for the analysis of NAEP data. *Journal of Educational Statistics*, 14(4), 303–334.

Kim, S., & Lee, W.-C. (2004). IRT scale linking methods for mixed-format tests. Iowa City, IA: ACT.

Klentschy, M. P. (2006). Science education in a No Child Left Behind, standards-based world. In R. Douglas, M. P. Klentschy, K. Worth, & W. Binder (Eds.), Linking science and literacy in the K–8 classroom (pp. 377–389). Washington DC: National Science Teachers Association Press.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement*, 11(3), 263–277.

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

Kosko, K., & Wilkins, J. (2011). Communicating quantitative literacy: An examination of open-ended assessment items in TIMSS, NALS, IALS, and PISA. Numeracy, 4(2), Article 3. http://doi.org/10.5038/1936-4660.4.2.3

Lamprianou, I. (2010). The practical application of optimal appropriateness measurement on empirical data using Rasch models. *Journal of Applied Measurement*, 11(4), 409–423.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4(4), 269–290.

Martin, M. O., & Mullis, I. V. S. (2012). Performance at the TIMSS 2011 international benchmarks. In M. O. Martin, & I. V. S. Mullis (Eds.), Methods and procedures in TIMSS and PIRLS 2011 (pp. 86–138) Chestnut Hill, MA: Boston College. http://timss.bc.edu/timss2011/downloads/T11_IR_MChapter2.pdf

Martin, M. O., Mullis, I. V. S., Foy, P., Brossman, B., & Stanco, G. M. (2012). Estimating linking error in PIRLS. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 5, 35–48.

Michaelides, M. P., & Haertel, E. H. (2004). Sampling of common items: An unrecognized source of error in test equating (CSE Report 636). Los Angeles, CA: Center for Research on Evaluation Standards and Student Testing (CRESST).

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, 8(3), 323–325.

Monseur, C., Sibberns, H., & Hastedt, D. (2008). Linking errors in trend estimation for international surveys in education. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 1, 113–122.

Mullis, I. V. S., Drucker, K. T., Preuschoff, C., Arora, A., & Stanco, G. M. (2012). Assessment framework and instrument development. In M. O. Martin, & I. V. S. Mullis (Eds.), Methods and procedures in TIMSS and PIRLS 2011. Chestnut Hill, MA: Boston College. http://timssandpirls.bc.edu/methods/t-instrument.html

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). TIMSS 2011 assessment frameworks. Chestnut Hill, MA: Boston College.

Muraki, E., & Bock, R. D. (2003). PARSCALE 4: IRT item analysis and test scoring for rating-scale data (Version 4) [computer software]. Chicago, IL: Scientific Software.

Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement*, 25(1), 53–67.

Organisation for Economic Co-operation and Development (OECD). (2012). PISA 2009 technical report. Paris, France: OECD Publishing.

Sheehan, K. M., & Mislevy, R. J. (1988). Some consequences of the uncertainty in IRT linking procedures. ETS Research Report Series, 1988(2), i–40.

Xu, X., & Davier, M. (2010). Linking errors in trend estimation in large-scale surveys: A case study. ETS Research Report Series, 2010(1), i–12.doi:10.1002/j.2333-8504.2010.tb02217.x

Zonts, J. M. (2013). The United States growth over 16 years of student correct responses on the TIMSS: Are we really that far behind? All Theses and Dissertations, Brigham Young University, Paper 3730.

Citation:

Hastedt, Dirk, & Desa, Deana (2015). Linking Errors Between Two Populations and Tests: A Case Study in International Surveys in Education. *Practical Assessment, Research & Evaluation*, 20(14). Available online: http://pareonline.net/getvn.asp?v=20&n=14

Corresponding Author:

Deana Desa
IEA Data Processing and Research Center
Mexikoring 37
22297 Hamburg
Germany

email: deana.desa [at] iea-dpc.de