Linking errors in trend estimation for international surveys in education

C. Monseur
University of Liège, Liège, Belgium

H. Sibberns and D. Hastedt
IEA Data Processing and Research Center, Hamburg, Germany

For a decade, more or less, one of the major objectives of international surveys in education has been to report trends in achievement. For that purpose, a subset of items from previous data collections has been included in new assessment instruments. The linking process (i.e., reporting the cognitive data from different data collections on a single scale) is implemented through item response theory (IRT) models. Under IRT assumptions, the same linking function is obtained regardless of which common items are used because item-specific properties are fully accounted for by the item's IRT parameters. However, model misspecifications always occur, such as small changes in the items, position effects, and curriculum effects. Therefore, other sets of linked items can generate other linking transformations, even with very large examinee samples. According to Michaelides and Haertel (2004), error due to the common-item sampling does not depend on the size of the examinee sample, but rather on the number of common items used. As such, the selection of anchor items may constitute the dominant source of error for summary scores. During its history, the International Association for the Evaluation of Educational Achievement (IEA) has reported trends in achievement for TIMSS 1999, TIMSS 2003, and PIRLS 2001, but has not accounted for linking errors in addition to the usual sampling and imputation errors, a situation that leads to an increase in Type I errors. It is for this reason that this study analyzes the variability of the trends estimate due to the selection and length of the anchor test used to link the assessments.

INTRODUCTION

The interest taken by policy-makers in monitoring education systems and measuring the effects of educational reforms has contributed to an increased emphasis on trend indicators in the design of recent surveys of educational achievement. Trends over time provide policy-makers with information not only on how the achievement levels of students in their country change in comparison with the achievement levels of students in other countries, but also on how within-country differences, such as gender gaps in achievement, evolve over time. The increasing emphasis on trend indicators has constituted a major change in international surveys of education over the past decade. The names of two current IEA surveys reflect this growing interest: the *Trends* in International Mathematics and Science Study (TIMSS), and the *Progress* in International Reading Literacy Study (PIRLS).

Under IRT assumptions, the same linking function should be obtained regardless of which common items are used because item-specific properties are fully accounted for by the item's IRT parameters. However, model misspecifications always occur, and no model fits real data perfectly. Factors contributing to misfit include small changes in the items, position effects, test design, and curriculum effects. This misfit means other sets of linked items can generate other linking transformations, even with very large examinee samples. According to Michaelides and Haertel (2004), error due to common-item sampling depends not on the size of the examinee sample but on the number of common items used. As such, the error due to the common-item sampling could constitute the dominant source of error for summary scores.

Although IEA reports trends indicators for achievement in its current studies, the association bases the standard error reported for the trends estimates only on the standard errors associated with the two mean achievement estimates used to compute the trends. This trend standard error estimate has two components—the sampling uncertainty and the measurement uncertainty. It therefore consists of (i) sampling variance and (ii) uncertainty about student performance, and it is reflected through the variance of the plausible values. In contrast, the PISA 2003 initial report, which also reports trends indicators in reading, adds another source of variance. As described in the PISA 2003 technical report (Organisation for Economic Co-operation and Development/OECD, 2005), the standard error on the trends estimates contains a third error component, denoted as the linking error. This error reflects model misfit, such as item parameter drift, between the two data collections. However, the linking error, as used in PISA 2003, appears to be unsatisfactory because:

1. It assumes item independency, which is inconsistent with the embedded structure of items into units (passages or blocks of items);
2. It requires that partial credit items be considered dichotomous items; and
3. It takes only the international misspecifications between the two data collections into account.

This situation can lead to researchers underestimating the linking errors and thus increasing the Type I error. This situation, in turn, results in researchers reporting a significant change in achievement when, in fact, the change may not be significant. Furthermore, researchers generally interpret and publish results without regard to the test used. In other words, IEA reports achievement results in terms of reading literacy, mathematics, and science in general and not in terms of, for example, reading literacy on a specific test, such as with the PIRLS test. It also appears to interpret an achievement trend in terms of change in the student performance and not in terms of change in achievement on the anchoring items. In this context, the political importance of trends in achievement should not be underestimated. Also, if scholars suggest educational reforms based on the significant shifts, they may actually end up offering inappropriate policy recommendations.

Throughout the history of international surveys of achievement in education, the IEA Reading Literacy Study has offered a unique opportunity to study the linking error. This is because the achievement test used in 2003 is exactly the same as the achievement test used in the IEA Reading Literacy Study of 1991. In other surveys, instruments differ, changes in the test design occur, and/or (as is the case in PISA) the relative importance of the domains vary from one data collection to another.

METHOD

Nine countries participated in both the IEA Reading Literacy Study 1991 and the Reading Literacy Repeat Study 2001. However, the data from only eight countries were reanalyzed (Greece, Hungary, Iceland, Italy, New Zealand, Slovenia, Sweden, and the United States). It was not possible to include the data from Singapore because these were unavailable at the time of analysis.

The Reading Literacy Study 1991 performance instrument consisted of 106 items administered to all students, without any rotation (Wolf, 1995). The first 40 items, which assessed "*word recognition*," were not included in our study.

> The Word Recognition part was followed by a number of reading passages and documents, for each of which a set of items were asked. Four reading passages with 22 items were selected from the expository domain, five passages with 21 items were selected from the expository domain, and six documents with 23 items were selected from the documents domain. (Elley, 1994, p. 10)

Two of the 66 items were deleted because they had been recoded "*not applicable*" for all students in a country.[1] We therefore had a pool of 64 items from which we could randomly select particular numbers of items.

We decided not to pursue alternatives, such as resampling methods based on the jackknife procedure, because the main focus of our study was (i) to empirically demonstrate the existence of a linking error, (ii) to analyze the significance of reporting

[1] The original test consisted of 68 items but two were deleted at the international level.

a common linking error for all countries, and (iii) to show the effect of anchor-test length on the linking error. However, the method of randomly selecting items from the pool that we adopted in this paper did not take into account the embedded structure of the test, that is, the set of items related to a single reading passage. Further, because we selected items from a finite pool of 64 items, the empirical linking error automatically became 0 when the number of selected concurrently calibrated items was equal to the whole set of items, namely the 64 in the case under consideration.

Let us suppose, then, that 20 items of the 64 were used in the IEA Reading Literacy 2001 study. This would have resulted in about 28 millions of billions of possible different tests from the 20 items out of the pool of 64 items. For this study, we constructed 50 tests of 20 items randomly selected from the item pool. We used the same method to construct 50 tests of 30 items, 50 tests of 40 items, and 50 tests of 50 items. We used ConQuest (Wu, Adams, & Wilson, 1997) to analyze each data set (i.e., eight countries by two data collections by 50 tests by four types of tests, or 3,200 data sets) and thereby draw plausible values.

Note that we did not use conditioning variables. The absence of conditioning enlarges the variance of the *posterior* distributions and therefore slightly increases the imputation error. It also underestimates the relationship between contextual variables and performance. However, because we mainly analyzed the difference in the country mean estimates between two data collections, we determined that any bias introduced through the absence of conditioning was acceptable given the additional computation time that a more sophisticated model would necessitate.

Before generating the plausible values, we drew random samples of 500 students per country and per data collection, and performed a joint calibration of the whole item pool so as to obtain the item parameters according to a one-parameter IRT model. We then transformed the plausible values on the *logit* scale on a new scale with a mean of 500 and a standard deviation of 100 by using *senate* weight per test, whatever the number of items included in the test. Thus, the distribution of the eight countries and the two data collections had a mean of 500 and a standard deviation of 100. We then computed the achievement trend per test by comparing the country mean at Time 1 (1991) and the country mean at Time 2 (2001). Finally, we computed the mean and the standard deviation of the 50 trends estimated for each type of test.

RESULTS

The average trends per type of test all correlated at 0.97 and are reported in the international report (Martin, Mullis, Gonzalez, & Kennedy, 2003). We could not expect a perfect correlation because Singapore was not included in the analyses. Also, the scaling model in this approach (1PL) differed from the model used in the 10-year trend study (3PL).

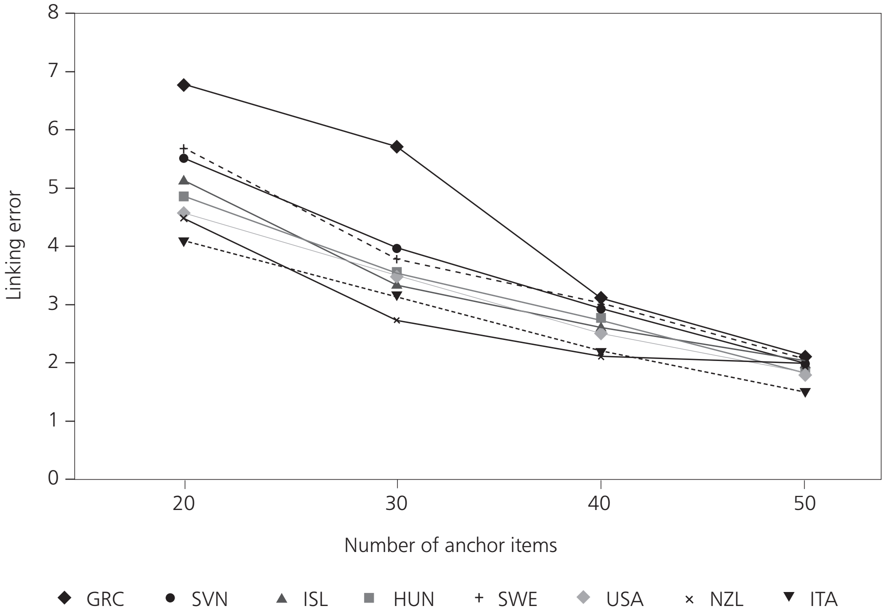
[2] Here, the sum of the student weights per country and per data collection is a constant, which means that each country contributed equally to the linear transformation.

Table 1 and Figure 1 present the linking error, that is, the standard deviation across the 50 trends estimates per type of test. As the table and figure show, the trend estimate for a particular country varies according to the selection of anchor items. For example, with tests of 20 items, the trends estimates for Greece range from 22 to 52. These results clearly demonstrate the impact of the item selection on the trend estimates and advocate the use of a linking error for testing the significance level of a particular trend. Because, in international surveys, the link between two data collections usually is based on fewer than 40 items, the linking error is quite substantial, and it is more or less the same size as the sampling error. For instance, the standard errors on the achievement trend estimates in PIRLS 2001 (Martin et al., 2003) ranged from 3.7 to 7.4. No doubt, the outcomes of the test would differ for countries with low trend estimates.

Table 1: Linking error (i.e., standard deviation of the 50 trend estimates) per country and per type of test

	GRC	HUN	ISL	ITA	NZL	SVN	SWE	USA
Test of 20 items	6.78	4.88	5.16	4.11	4.51	5.52	5.64	4.6
Test of 30 items	5.74	3.57	3.41	3.24	2.79	4.00	3.83	3.54
Test of 40 items	3.15	2.76	2.67	2.21	2.13	2.97	3.07	2.56
Test of 50 items	2.15	1.85	2.05	1.53	2.00	2.00	2.08	1.84

Figure 1: Linking error per country and per type of test



How do these empirical linking errors compare with the analytic solution adopted in the PISA 2000 technical report (Adams & Wu, 2002)? Here, we computed the variability of the shift in item parameters between their 1991 estimates and their 2001 estimates and then transformed them on the IEA Reading Literacy scale. On using Formula 1 (below), we found the linking error was equal to 4.2 for a test with 20 anchor items, to 3.4 for a test with 30 anchor items, to 3.0 for a test with 40 anchor items, and to 2.7 for a test with 50 anchor items.

$$\sigma_{(linking)} = \sqrt{\frac{\sigma^2_{(shift)}}{n_{(anchor)}}} \qquad (1)$$

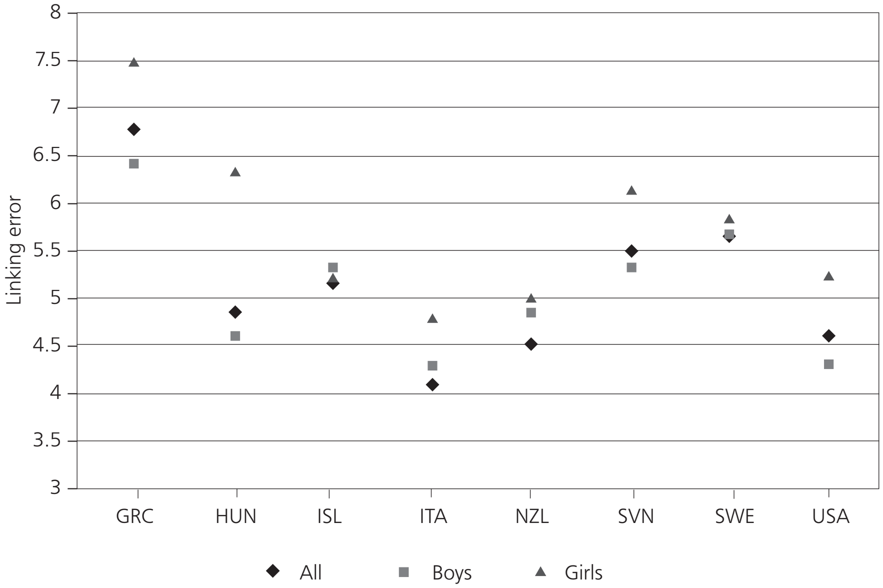
Because the method we adopted in this paper assumes a finite population of items, and because Formula 1 assumes an infinite population of items, the empirical linking error estimate and the analytical linking error estimate do not converge as the number of anchor items increases. However, the inconsistencies between the two estimates for a small number of items are noticeable. The analytical solution apparently underestimates the linking error. Table 1 and Figure 1 also show the variability of the linking error from one country to another for a particular test type. For example, the linking error is 6.78 for Greece but only 4.11 for Italy. This observation implies that a single linking error for all countries was not as accurate as it should have been.

Why do some countries present a larger linking error? As shown in Table 2, the size of the linking error correlates highly with the importance of the trends estimates. We can expect this observation to some extent because the posterior variance is greatest at the extremes, but this is not the case if the test is targeted to the most proficient or the least proficient populations. Furthermore, if the linking error varies at the country level, we could expect the linking error to also vary across sub-populations within countries. As an illustration, Figure 2 presents the linking error per country and per gender for a test of 20 items.

Table 2: Correlation between the trend estimate (expressed in absolute value) and its linking error

Type of test	Correlation
20	0.91
30	0.88
40	0.82
50	0.66

Figure 2: Overall linking error for 20-item linking and linking error by gender



Our analysis thus far has identified two factors influencing the size of the linking error. The first is the number of items and the second is the size of the trends. But we also need to consider two other factors—the embedded structure of the items and the modification in the test design.

Most of the PISA and the IEA Reading Literacy and PIRLS assessment materials present a hierarchical structure: items are clustered in units. A unit consists of a stimulus, that is, a reading passage in the case of the Reading Literacy assessment and a contextualization for the PISA mathematics and science literacy assessment, followed by a set of items all related to that stimulus. By adapting Formula 1 to a cluster sample, we obtain, in the case of a constant number of items per unit, the following:

$$\sigma_{(linking)} = \sqrt{\frac{\sigma^2_{(between\_unit\_shift)}}{n_{(anchor\_unit)}} + \frac{\sigma^2_{(within\_unit\_shift)}}{n_{(anchor\_unit)}}}$$

Monseur and Berezner (2006) reported a substantial increase of the linking error with items organized in units. These authors also have analyzed, through simulations, the accuracy of a jackknifing method and analytical solution for estimating the linking error with hierarchically structured items. The analytical solution and the jackknife methods provide estimates that do not significantly differ from the empirical estimates of the linking error.

A change in the test design constitutes the second additional factor that can affect the size of the linking error. The IEA Reading Literacy study is of particular interest in this regard, as no changes were made in the test instruments. We used a variance decomposition with three factors—country, item, and time—to analyze the national item parameters from a one-parameter IRT model. Because the items are centered for any country at any time, the time variance, the country variance, and the time-by-country interaction variance are equal to 0.

We also conducted this analysis on the PISA anchor reading items between the 2000 and the 2003 data collections. The PISA 2000 test design consisted of nine tests, with four blocks of items for each test. The 28 anchor items appeared only in the first three blocks. In 2003, these 28 items were distributed into two clusters of reading items and appeared once in each of the four positions.

Table 3 presents the estimation of the variance components. Here we can see that in the IEA Reading Literacy study, the time-by-item interaction is about one third of the time-by-item-by-country interaction. However, in PISA, the time-by-item interaction is about twice the value of the time-by-item-by-country interaction. A modification in the test design can therefore have marked consequences for the size of the linking error. This last observation should encourage test developers of international surveys in education to avoid, or at least minimize, changes in the test design between two data collections.

Table 3: Variance decomposition of the national item parameter

Source of variation	IEA Reading Literacy	OECD PISA
Item	1.02590	0.95443
Country by item	0.17701	0.14794
Time by item	0.01083	0.04040
Time by item by country	0.03090	0.02758

CONCLUSION

In 2004, the PISA 2003 initial report published by the Organisation for Economic Co-operation and Development (OECD, 2004) reported trends. As described in the OECD PISA 2003 technical report (OECD, 2005), the standard error of the trend estimate included a linking error. However, as Monseur and Berezner (2006) pointed out, the addition of a linking component in the standard error in the study constituted a methodological improvement but did raise several issues. Essentially, the linking error as used in PISA 2003 seemed unsatisfactory for the same reasons as those outlined in the introduction to this paper.

The results of the simulations presented in this study highlight the relationship between the number of items and the linking error and (more importantly) the variability of the linking error from one country to another. The linking error also correlated highly with the achievement trend estimates. The results additionally highlight the increase in the linking error for within-country analyses as shown by the gender example. Finally, the analyses presented in this paper outline the danger of modifying the test design on the linking error.

Further analyses should now be devoted to computing the linking error on the final set of anchoring items. Replication methods like jackknifing and bootstrapping usually used in the sampling area might be of interest. While an analytical solution might be adopted for simple contexts, jackknifing presents no restriction. It can be used with two- or three-parameter IRT models, with polytomous items, and with hierarchically structured items where units do not necessarily have the same number of items.

If policy-makers and international report readers limited their interpretation of the trend estimates to the anchoring items, it would not be necessary to recommend the addition of a linking error. However, an improvement in student performance based on several dozen anchor-items is currently interpreted by researchers and policy-makers as an improvement in student performance for the whole domain assessed by the study. As such, the inclusion of a linking error in reporting trends would be consistent with how trends are presently interpreted.

According to Michaelides and Haertel (2004), common items should be considered as chosen from a hypothetical infinite pool of potential items. Cronbach, Linn, Brennan, and Haertel (1997) also adhere to this point of view. Remember that a test score is based on an examinee's performance on a particular test form consisting of certain items. What is therefore of most interest is not how well the examinee did on those particular items at that particular occasion. Rather, it is the inference drawn from that example of performance to what the examinee could do across many other tasks requiring the application of the same skills and knowledge.

The interpretations of the trends indicators by policy-makers and the arguments presented by scholars like Michaelides and Haertel and Cronbach and colleagues advocate hypothetical infinite populations. In other words, even if a new international test did include all items from a previous survey, a linking error would still need to be reported. This linking error would reflect the model misspecifications.

Limitations of the current research are that we used only two cycles of a reading assessment, and that we could compute only linking errors for true subsets of the anchor test. However, the research presented here demonstrates that linking error is a potential source of variation that can be quantified through computational procedures involving resampling methods, such as the bootstrap and the jackknife (Efron, 1982). Future research could focus on methods that take linking error as well as sampling error into account simultaneously. One way of doing this is outlined in Cohen, Johnson, and Angeles (2001). However, their approach to jackknifing in two dimensions needs careful examination in terms of whether it is executed correctly and yields appropriate variance estimates. But whether executed as identifying separable sources, or whether carried out simultaneously, research that incorporates additional tractable sources of variation promises to improve comparisons both across and within countries for subgroups of interest to policy-makers and educators.

References

Adams, R. J., & Wu, M. (Eds.). (2002). *PISA 2000 technical report*. Paris: Organisation for Economic Co-operation and Development.

Cohen, J., Johnson, E., & Angeles, J. (2001). *Estimates of the precision of estimates from NAEP using a two-dimensional jackknife procedure*. Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*, 373–399.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, No. 38.

Elley, W. B. (1994). *The IEA study of reading literacy: Achievement and instruction in thirty-two school systems*. London: Pergamon.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS trends in children's reading literacy achievement 1991–2001*. Chestnut Hill, MA: Boston College.

Michaelides, M. P., & Haertel, E. H. (2004). *Sampling of common items: An unrecognized source of error in test linking*. Los Angeles: Center for the Study of Evaluation (CSE), University of California.

Monseur, C., & Berezner, A. (2006). *The computation of linking error*. Paper presented at the AERA annual convention's symposium on measuring trends in international comparative research: Results from the first two cycles of the OECD/PISA study, San Francisco.

Organisation for Economic Co-operation and Development (OECD). (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: Author.

Organisation for Economic Co-operation and Development (OECD). (2005). *PISA 2003 technical report*. Paris: Author.

Wolf, R. M. (1995). *The IEA Reading Literacy study: Technical report*. The Hague: International Association for the Evaluation of Educational Achievement.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest: Multi-aspect test software* (computer program). Camberwell, VIC: Australian Council for Educational Research.