Equating errors in international surveys in education

C. Monseur
University of Liège
Liège, Belgium
H. Sibberns
IEA Data Processing Center
Hamburg, Germany
D. Hastedt
IEA Data Processing Center
Hamburg, Germany

Abstract

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 a new assessment test and the equating process (i.e., reporting the cognitive data of different data collections into a single scale) implemented through Item Response Theory models. Under IRT assumptions, the same equating 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 equating 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, this error could 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 through TIMSS 1999, TIMSS 2003, and PIRLS 2001, but has not added equating errors to the usual sampling and imputation errors, leading to an increase in Type I errors. It is for this reason that this study analyzes the variability of the trends estimate.

Introduction

Policymakers' interest in the monitoring of education systems and in 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 policymakers with information not only on how the achievement level of students in their country changes 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 progressive emphasis on trend indicators constitutes 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 equating 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. These include small changes in the items, position effects, and curriculum effects. This means other sets of linked items can generate other equating 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, common error due to the common-item selection could constitute the dominant source of error for summary scores.

Although IEA reports trends indicators for achievement in its current studies, it bases the standard error on the trends estimates only on the standard

errors associated with the two mean achievement estimates used to compute the trends. The standard error has two components: the sampling uncertainty and the measurement uncertainty.[1] The PISA 2003 initial report also reports trends indicators in reading. As described in the PISA 2003 technical report, the standard error on the trends estimates adds a third error component, denoted linking error. This error reflects model misspecifications between the two data collections. However, the PISA 2003 linking error appears to be unsatisfactory because:

1. It assumes item independency, which is inconsistent with the embedded structure of items into units;
2. It considers partial credit items as dichotomous items; and
3. It takes only the international misspecifications between the two data collections into account.

Not recognizing the uncertainty due to the linking process leads to an underestimation of the linking errors and thus increases the Type I error, thereby resulting in the reporting of significant changes in achievement when, in fact, these are not significant. Furthermore, results are usually interpreted and published 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 is also very likely that an achievement trend will be interpreted in terms of change in the student performance and not in terms of changes 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 equating error. This is because the achievement test used in 2003 is exactly the same achievement test used by the IEA Reading Literacy Study in 1991. Indeed, in other surveys, instruments are different, changes in the test design can occur, or, as is the case in PISA, the relative importance of the domains can vary from one data collection to another.

Method

Nine countries participated in both the IEA Reading Literacy Study of 1991 and the Reading Literacy Repeat Study of 2001. However, the data from only eight countries were re-analyzed (Greece, Hungary, Iceland, Italy, New Zealand, Slovenia, Sweden, and the United States). For timing reasons, it was not possible to include the Singapore data.

The Reading Literacy Study 1991 performance instrument consisted of 108 items administered to all students, without any rotation (Wolf, 1995). The first 40 items, which assessed "*word recognition*," were not included in our study because they showed a severe ceiling effect. Of the 68 remaining items, we deleted three from the database because they had been recoded "*not applicable*" for all students. We therefore had a pool of 65 items from which we could randomly select particular numbers of items.

Simulations were used to empirically compute the linking error. Two factors that might have had an impact on the linking distribution for the IEA Reading Literacy Study were (i) the number of anchor items, and (ii) the importance of the shifts in the item parameters between the two data collections. This present study therefore analyzed the variability of the linking error depending on the number of anchor items by using replication methods. Let us suppose that 20 items of the 65 were used in the IEA Reading Literacy Repeat Study. This would have resulted in about 28 millions of billions of possible tests of 20 items out of the pool of 65.

For this study, 50 tests of 20 items randomly selected from the item pool were constructed. The same method was used to construct 50 tests of 30 items, 50 tests of 40 items, and 50 tests of 50 items. Each of the data sets (i.e., eight countries by two data collections by 50 tests by four types of tests or 3,200 data sets) was submitted to ConQuest (Wu, Adams, & Wilson, 1997) for drawing plausible values. Note that no conditioning variable was used.

Before generation of the plausible value, random samples of 500 students per country and per data collection were drawn, and a joint calibration of the whole item pool performed to obtain the item parameters. The plausible values on the *logit* scale were

1 Because student performance estimates are reported through plausible values, the measurement uncertainty corresponds to the imputation variance.

then transformed on a new scale with a mean of 500 and a standard deviation of 100 by using *senat* weight per test,[2] 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. The achievement trend was then computed per test by comparing the country mean at Time 1 (1991) and the country mean at Time 2 (2001). Finally, the mean and the standard deviation of the 50 trends estimated were computed per type of test.

Results

The average trends of the test all correlated with the reported trends in the international report (Martin, Mullis, Gonzalez, & Kennedy, 2003). A perfect correlation could not be expected because one country was not included in the analyses. Also, the scaling model in this approach (1PL) was different from the model used in the 10-year trend study (3PL).

Table 1 and Figure 1 present the linking error, that is, the standard deviation across the 50 trends estimate per type of test. As the table and figure show, the variability of the trends increases as the number of items decreases. 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, as it has more or less the same size as the sampling error. For instance, the standard errors on the achievement trend estimates in PIRLS Repeat (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.

Figure 1: Linking Error per Country and per Type of Test

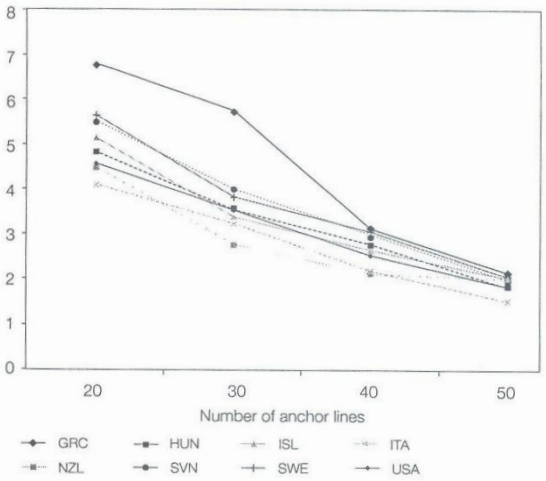


Table 1 and Figure 1 also show the variability of the linking error from one country to another for a particular test type. This observation implies that a single linking error for all countries is not as accurate as it should be. For example, the linking error is 6.78 for Greece but only 4.11 for Italy. Different analyses therefore were implemented in order to understand the outlying linking error for Greece.

First, the variability of the shifts in the national item parameters between 1991 and 2001 was computed for the eight countries. As expected, the variance of the national item parameters correlated at 0.49 with the linking error. In other words, the larger the shifts in the national item parameters, the larger the linking error. However, the factor that seemed to contribute mainly to the size of the linking error was the trends estimate. Table 2 provides the correlation between the *absolute* value of the trend estimates and the linking

Table 1: Linking Error 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	4.60
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

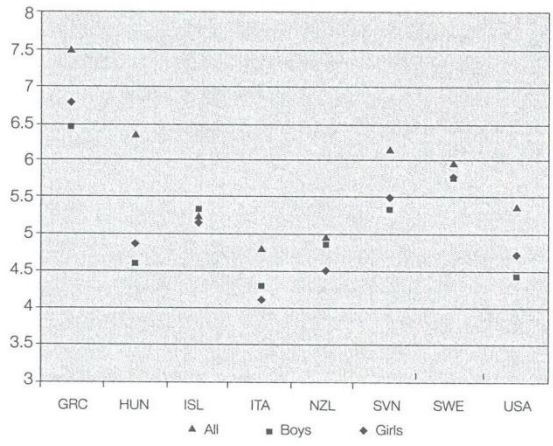
2 Here, the sum of the student weights per country and per data collection is a constant, which means that each country contributed equally to the linear transformation.

errors per type of test. The table shows that as the trend estimate increased, the linking error increased. Finally, the linking error was computed for each country and gender. Table 3 and Figure 2 present the overall linking error, as well as the linking errors for gender. In three countries (Iceland, New Zealand, and Sweden), there was nearly no difference between the overall linking error and the linking error for each gender. For Hungary, the linking error for girls was actually higher than the overall linking error. Note, however, that for all countries, the linking error was higher for girls than for boys. Further research is necessary to explain these differences. It is possible that the item format was the main cause for the observed differences.

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 and Linking Error per Gender



Conclusion

In 2004, the OECD PISA 2003 initial report (Organisation for Economic Co-operation and Development/OECD, 2004) also reported trends. However, as described in the OECD PISA 2003 technical report (OECD, 2005), the standard error of the trend estimate included a linking error. As discussed in Monseur and Berezner (2006), while the addition of a linking component in the standard error constituted a methodological improvement, it did raise several issues. In particular, the PISA 2003 linking error appears to be unsatisfactory because:

1. It made the assumption of item independency, which is inconsistent with the embedded structure of items into units;
2. It considered partial credit items as dichotomous items; and
3. It took into account only international misspecifications between the two data collections.

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 also highlight the increase of the linking error for within-country analyses as shown by the gender example.

Further analyses should now be devoted to computation of 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.

If policymakers 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 as an improvement for the students for the whole

Table 3: Overall Linking Error and Linking Error for Gender

	GRC	HUN	ISL	ITA	NZL	SVN	SWE	USA
All	6.78	4.88	5.16	4.11	4.51	5.52	5.64	4.60
Boys	6.43	4.60	5.33	4.29	4.85	5.32	5.63	4.30
Girls	7.50	6.36	5.24	4.82	4.98	6.15	5.86	5.26

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 being 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 of most interest here 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 policymakers and the arguments presented by scholars like Michaelides and Haertel (2004) and Cronbach et al. (1997) advocate for 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.

References

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, 5*(7), 373–399.

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 equating.* 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.