Estimating linking error in PIRLS

Michael O. Martin, Ina V. S. Mullis, Pierre Foy, Bradley Brossman, and Gabrielle M. Stanco

Boston College, Chestnut Hill, Massachusetts, United States

The Trends in Mathematics and Science Study (TIMSS) and Progress in Reading Literacy Study (PIRLS), as well as other large-scale assessments, measure changes in student achievement over time by linking one assessment to the next. Linking error is 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 cycle to the next. The estimation of the scale-linking transformation is, as with any statistical approximation, susceptible to estimation error. This study describes the method used to estimate linking error for the TIMSS and PIRLS assessments and examines the magnitude of linking error between the PIRLS 2001 and 2006 assessments. As anticipated, linking error was small and had little impact on significance tests of achievement differences between the two assessments, most likely because almost half the items were common to both PIRLS assessments.

IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments Volume 5

INTRODUCTION

Policymakers have become increasingly interested in student achievement trends that provide information about changing patterns of student achievement and enable them to monitor the results of educational reforms over time. To measure changes in student achievement, trend assessments operate on a regular cycle, administering a pool of achievement items to comparable samples of students every three to five years. IEA's Trends in International Mathematics and Science Study (TIMSS) has been measuring trends for some 70 participating countries every four years since 1995 (i.e., 1999, 2003, 2007, and 2011). IEA's Progress in International Reading Literacy Study (PIRLS) is conducted every five years, with assessments taking place in 2001, 2006, and 2011.

Because of the necessity for public disclosure and the need to ensure the current relevance of each new assessment, a selection of items from the assessment is released after each assessment cycle and replaced by newly developed items in the next. To maintain comparability across cycles, however, it is also necessary to have a substantial number of items that are not released and that are included in adjacent cycles. With the common items from successive assessments used as the basis for linking, scores from each new assessment cycle are then placed on the existing achievement scale from previous assessment cycles, so allowing differences from one assessment cycle to the next to be measured. The fact that some items are released and replaced by others means that the assessment changes somewhat from cycle to cycle, which introduces linking error.[1]

In large-scale assessments such as TIMSS and PIRLS, linking error is the error associated with placing achievement data from the most recent assessment cycle on a preexisting trend scale. As such, this error has two major sources:

1. Changes in the pool of items used to measure achievement as previously used items are released and replaced by new items; and
2. Shifts in the measurement properties of the common items from one assessment cycle to the next.

An effective linking method ensures that linking error is kept to a minimum while providing an estimate of the magnitude of the error and its impact on estimates of change in student achievement from cycle to cycle. In TIMSS and PIRLS, new replacement items are developed according to the same assessment frameworks as the released items, thereby ensuring that they have a similar focus and subject-matter coverage. In particular, the new items address mathematics and science content and cognitive domains that are the same as those for TIMSS and reading purposes and comprehension processes that are the same as those for PIRLS. Furthermore, in order to provide sufficient number of common items to maintain a stable link across assessment cycles, the TIMSS and PIRLS assessment designs currently specify that each assessment

[1] Note that linking error applies only when comparing achievement results across cycles. Linking error is not an issue when making comparisons within the same assessment cycle.

shares 60% of its items with the next assessment cycle. For the TIMSS eighth-grade assessments, this specification resulted in 126 items common to the 2007 and 2011 mathematics assessments, and 125 items common to the science assessments.

Due to the increased visibility of and reliance on trend data, examining the precision of estimates of trends in student achievement and the error associated with these estimates has become an important area of research. This study describes the TIMSS and PIRLS approach to measuring student achievement trends, and presents a method for estimating linking error based on this approach. The method is then applied to estimate linking error between the PIRLS 2001 and 2006 assessments.

IRT APPROACHES TO MEASURING TRENDS

Large-scale assessments of student achievement, such as the National Assessment of Educational Progress (NAEP) in the U.S. and the TIMSS, PIRLS, and Program in International Student Assessment (PISA) international assessments, rely on item response theory (IRT) methods to construct achievement scales for reporting student achievement and measuring trends from assessment cycle to assessment cycle. IRT methods are valuable in this context because they provide a way to estimate achievement in a student population based on the measurement properties of the individual items comprising the assessment. The item properties, or item parameters, are not known in advance, but must first be estimated from the assessment data through a process known as item calibration.

On completion of the item calibration, the item parameters are used to produce estimates of student achievement, which in large-scale assessments are typically in the form of "plausible values." These are estimates of student performance on the entire assessment, conditional on the responses the students gave to the assessment items they were administered and on the students' background characteristics (Foy, Galia, & Li, 2007).

Typically, one of two methods is used to calibrate item parameters within the IRT framework. In the separate calibration method, assessment data from adjacent cycles are calibrated separately (Kolen & Brennan, 2004). That is, the assessment data for the previous cycle are calibrated first, after which the assessment data for the current cycle are calibrated. A scale linking method is then used to place parameter estimates from the two calibrations on the same scale.

In contrast to the separate calibration method, the concurrent calibration method uses all data from both the previous cycle and the current cycle to estimate item and person parameters at the same time (Kolen & Brennan, 2004). Given that all parameters are estimated at the same time and therefore items common to both cycles receive the same estimates, item parameters from both cycles are on the same scale when the concurrent calibration method is applied. An advantage of the concurrent calibration approach is that it makes maximum use of all available data to estimate the item parameters. Also, by recalibrating the item parameters for common items at each assessment cycle, it permits these parameters to evolve gradually across successive cycles as circumstances change.

After completion of the item calibration, student achievement is estimated using the newly calibrated item parameters. In large-scale assessments such as TIMSS and PIRLS, this procedure involves conducting a principal components analysis (PCA) using background variables and then developing a regression equation using both the principal component variables and the item responses to estimate plausible values for each student—a process known as "conditioning." Student achievement is estimated by using each of the sets of plausible values; variation across the sets of plausible values reflects the measurement error (Foy et al., 2007).

Once student achievement has been estimated, the scale-linking transformation places the current-cycle data onto the previously existing trend scale. In large-scale trend studies, the scale-linking transformation is typically determined by matching characteristics of the achievement distribution of the data from the previous cycle (obtained using the new item calibration) to characteristics of the achievement distribution of the same data on the existing trend scale (obtained using the previous item calibration). After the linear transformation that best matches these two distributions has been determined, the transformation is applied to the current cycle data, so allowing these data to be placed on the trend scale (Donoghue & Mazzeo, 1992; Muraki, Hombo, & Lee, 2000).

PISA is a large-scale international trend study that has investigated scale linking error in regard to trend estimation. The PISA approach to trend estimation incorporates the separate calibration method using Rasch scaling, followed by plausible value estimation and scale linking performed by matching characteristics of the achievement distributions. Linking error is then calculated based on Rasch item parameter estimates for the subset of common or link items. Specifically, linking error in PISA is calculated as the standard error of the mean difference in the Rasch difficulty parameters of the common items calibrated from two adjacent cycles.

To calculate linking error between the 2000 and 2003 reading and science administrations, for example, PISA first estimated the Rasch difficulty parameters for the 2003 assessments. The common item parameters were then centered by setting the mean to 0. The Rasch difficulty parameters for the common items in the 2000 administration were also centered by setting the mean to 0. The difference between the two item parameters was calculated and averaged across all items, thus representing the difference in relative difficulty of the common items across cycles. Finally, linking error was computed as the standard error of the mean difference using the traditional formula for calculating the standard error of a mean (Monseur & Berezner, 2007; OECD, 2005). This approach assumes that the common items are a random sample of all possible common items.

It is evident that linking error in PISA is a function of both item drift (differences in item difficulty from one administration to the next) and the number of common items across surveys. The linking error formula was modified for PISA 2006 to account for the clustering of items within passages/blocks for the reading assessment as well as for the inclusion of partial credit items (OECD, 2009).

Trend estimates for the PISA reading and science literacy scales linking PISA 2000, PISA 2003, and PISA 2006 were based on 22 to 28 common items and had linking error estimates ranging from 3.11 to 5.30 points on the PISA (500, 100) achievement scales. Trends in mathematics literacy from 2003 to 2006 were based on 48 common items and had a somewhat smaller link error of 1.38 points.

MEASURING TRENDS IN TIMSS AND PIRLS

Similar to other large-scale studies, TIMSS and PIRLS use IRT-based scaling methodology to estimate item parameters, generate plausible values for each student (with these values based on the conditioning model), and then link the current assessment scale to the trend scale. To make maximum use of the data from successive assessment cycles, TIMSS and PIRLS use the concurrent calibration method, as opposed to the separate calibration method, to estimate item parameters.

In this approach, all assessment items from both the current assessment and the previous assessment are included in the calibration for all countries participating in both assessment cycles. The PIRLS 2001 assessment, for example, consisted of eight passages, four literary and four informational, with a total of 98 items (133 score points). In the next cycle of PIRLS (2006), half of the PIRLS 2001 assessment (four passages and 49 items, worth 66 points) was reassessed for the purpose of trend measurement. The PIRLS 2006 assessment accordingly consisted of 10 passages,[3] five literary and five informational, with a total of 125 items with 165 score points. Four of these passages (and 49 items) were trend passages from 2001, and the other six passages and item sets (three literary and three informational) were newly developed to reflect the current environment and context of reading literacy (Martin, Mullis, & Kennedy, 2007). In summary, 49 items were unique to 2001, 49 items were common to 2001 and 2006, and 76 items were unique to 2006 (Foy et al., 2007).

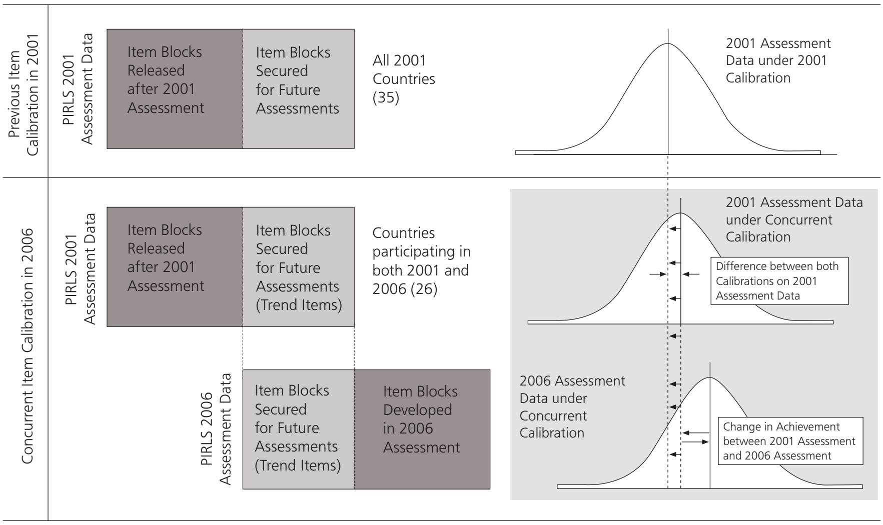
Figure 1 shows the concurrent calibration model used for PIRLS 2006. The right-hand side of the top panel shows that scaling the PIRLS 2001 achievement data resulted in a distribution of student results. That is, after item parameters were estimated in the PIRLS 2001 calibration, these item parameters were used to "score" the student responses and yielded the achievement results published in the *PIRLS 2001 International Report* (Mullis, Martin, Gonzalez, & Kennedy, 2003).

As explained in the *PIRLS 2001 Technical Report* (Martin, Mullis, & Kennedy, 2003), the process of estimating achievement used information about each student's responses to the administered items and the student's background characteristics to impute student scores, or plausible values, on the assessment as a whole. To quantify error in the imputation process, five plausible values were generated for each student, and all analyses were conducted five times. The average of the five analyses was taken as the

[3] In PIRLS 2006, the assessment was extended to include 10 passages instead of eight passages as in PIRLS 2001. PIRLS 2011 also includes 10 passages, of which six passages containing 75 items are common to both PIRLS 2006 and PIRLS 2011.

best estimate of the statistic in question, and the variance among them reflected the imputation error. The PIRLS achievement scale metric was established, based on the 2001 data, as having a mean of 500 and a standard deviation of 100.

Figure 1: Concurrent calibration model used for PIRLS 2006


Previous Item Calibration in 2001
PIRLS 2001 Assessment Data
Item Blocks Released after 2001 Assessment
Item Blocks Secured for Future Assessments
All 2001 Countries (35)
2001 Assessment Data under 2001 Calibration
Concurrent Item Calibration in 2006
PIRLS 2001 Assessment Data
Item Blocks Released after 2001 Assessment
Item Blocks Secured for Future Assessments (Trend Items)
Countries participating in both 2001 and 2006 (26)
2001 Assessment Data under Concurrent Calibration
Difference between both Calibrations on 2001 Assessment Data
PIRLS 2006 Assessment Data
Item Blocks Secured for Future Assessments (Trend Items)
Item Blocks Developed in 2006 Assessment
2006 Assessment Data under Concurrent Calibration
Change in Achievement between 2001 Assessment and 2006 Assessment

Source: IEA Trends in International Mathematics and Science Study (TIMSS) 2007.

Figure 1 also indicates (top left) that, after publication of the PIRLS 2001 achievement results, about half of the PIRLS 2001 items were released to the public and the other half were kept secure to be used again in PIRLS 2006. The lower panel of Figure 1 shows the concurrent calibration used in PIRLS 2006. The 2006 calibration included only those countries that participated in both 2001 and 2006. For these 26 countries, data from the entire PIRLS 2001 assessment were rescaled together with the new data from the PIRLS 2006 assessment.

To resolve the issue of metric indeterminacy inherent in IRT procedures, the scale was set such that the mean and standard deviation of the combined distribution were approximately 0 and 1, respectively. The item parameters from the 2006 concurrent calibration were used to score the 2006 data and produce the achievement results published in the *PIRLS 2006 International Report* (Mullis, Martin, Kennedy, & Foy, 2007).

As can be seen from the top panel of Figure 1, the PIRLS 2001 items had a set of item parameters from the original item calibration conducted in 2001. With the concurrent calibration approach, item parameters for the trend items are not fixed, but are re-estimated with each cycle. Thus, the PIRLS 2006 concurrent calibration, using all the data from 2001 and 2006 for countries that participated in both cycles, resulted in a second set of item parameters for the 2001 items. The second set of item parameters reflects changes in the pool of items used to measure achievement and shifts in the item parameter estimates. As part of this process, the two sets of parameters for the trend items were compared item by item, and the shifts typically were minor.

As shown in the lower panel of Figure 1, there was also a small change between the distribution of the PIRLS 2001 data under the original item calibration and the distribution of the PIRLS 2001 data under the concurrent calibration. The difference was due to changes in item parameters from the original item calibration in 2001 (based on 2001 data) and the concurrent calibration item parameters in 2006 (based on both 2001 and 2006 data).

PIRLS 2006 Scale Linking

The method used to place the PIRLS 2006 data on the trend scale is based on a procedure used in the National Assessment of Educational Progress (NAEP), described by Donoghue and Mazzeo (1992) and Muraki et al. (2000) and further investigated by Lee, Song, and Kim (2004). First, a linear transformation was performed to match the distribution of the 2001 data under the concurrent (2006) item calibration to the distribution of the 2001 data under the original (2001) item calibration. This transformation ensured that the mean and standard deviation of the distribution of 2001 data under the concurrent calibration aligned to the mean and standard deviation of the distribution of 2001 data from the original item calibration (i.e., the data were placed on the same scale). Next, the same linear transformation was applied to the PIRLS 2006 data. This placed the PIRLS 2006 data on the PIRLS 2001 metric, allowing the achievement trend between PIRLS 2001 and PIRLS 2006 to be estimated.

In summary, four steps were taken in PIRLS 2006 to estimate item parameters and ability distributions and to place these estimates on the trend scale:

1. Based on trend countries (i.e., countries that participated in both assessments), item parameters were estimated via concurrent calibration for all items on both the PIRLS 2001 and PIRLS 2006 assessments.
2. Achievement distributions were estimated for the PIRLS 2001 trend countries using the item parameters from the concurrent calibration.
3. The linear transformation was determined that best matched the PIRLS 2001 achievement distributions estimated under the concurrent calibration to the PIRLS 2001 achievement distributions published in 2001.
4. The linear transformation determined in (3) was applied to the PIRLS 2006 achievement distributions to place the current estimates on the trend scale.

Estimating Linking Error in PIRLS 2006

On completion of the concurrent calibration in 2006, each PIRLS 2001 item had two item parameter estimates: one estimate based on the 2001 calibration and one estimate based on the 2006 concurrent calibration. These made it possible to produce two achievement estimates from the 2001 data, one using the 2001 item parameters and the other using the 2006 parameters. Differences between these two estimates reflect linking error in PIRLS.

Linking error was computed as a function of the standard errors of the differences in achievement estimates resulting from the two sets of item parameters. First, plausible values for students in each trend country were generated using the 2001 student responses with the 2001 item parameters (i.e., the results published in the *PIRLS 2001 International Report*). All 2001 item parameters—as opposed to only common item parameters—were used to obtain achievement estimates so as to be consistent with the 2001 scaling and also to base achievement estimates on all available data. Plausible values were then generated for students in each trend country using the same data (i.e., the 2001 student responses), but substituting the reestimated (2006) item parameters for the original 2001 item parameters. Because the student responses were identical under both conditions, the difference between the two achievement estimates reflects the effect of the changes made to item parameters between 2001 and 2006, and its standard error may be considered an estimate of the linking error (Johnson, 2005).

The variability due to linking was estimated by jackknifing[4] the differences between the two achievement estimates across the five plausible values. Thus, student-level differences on the first plausible value were determined between estimates obtained on the 2001 data using the 2001 item parameters and estimates obtained on the 2001 data using the 2006 item parameters. These differences were then calculated for the second, third, fourth, and fifth plausible values. The average difference across each of the five plausible values was then calculated as the statistic of interest, and its standard error was estimated using the jackknife procedure.

The linking error was estimated separately for each country so that country-level results in PIRLS 2006 could be adjusted accordingly. Twenty-six countries and two Canadian provinces participated in both PIRLS 2001 and PIRLS 2006, and linking error was estimated for each participant. This approach differs from the approach utilized by PISA, which calculates only one overall linking error across all participating countries.

Linking Error Results for PIRLS 2006

Table 1 presents the PIRLS 2006 linking error results. The first two columns in the table present the average achievement and standard error for the PIRLS 2001 data based on the 2001 item parameters, which are the results originally reported in PIRLS

[4] See Wolter (2007) for a description of the jackknife procedure.

2001 (Mullis et al., 2003). The next two columns display the average achievement scores and the respective standard errors based on the same data using the item parameters from the 2006 concurrent calibration. The last two columns present the differences between average achievement based on the original 2001 calibration and average achievement based on the 2006 concurrent calibration. The standard errors of the differences are the linking-error estimates for the link between PIRLS 2001 and PIRLS 2006. The differences between achievement estimates are very small, mostly less than half a score point. The linking errors are also very small, ranging from 0.6 to 2.0, with an international average of 1.1.

Table 1: Average achievement and linking-error estimates using PIRLS 2001 data

Country	2001 data estimated from PIRLS 2001 calibration		2001 data estimated from PIRLS 2006		Difference in average achievement	
	Average achievement	*SE*	Average achievement	*SE*	Achievement difference	*SE* of difference (linking error)
Bulgaria	550	3.8	551	3.8	0.3	1.0
England	553	3.4	552	3.3	-0.6	1.4
France	525	2.4	526	2.4	0.4	0.7
Germany	539	1.9	539	1.8	0.0	0.8
Hong Kong SAR	528	3.1	528	3.2	0.3	0.8
Hungary	543	2.2	544	2.1	0.3	1.1
Iceland	512	1.2	512	1.2	-0.7	1.1
Iran, Islamic Rep. of	414	4.2	414	4.4	0.6	1.5
Israel	509	2.8	509	2.8	0.2	1.2
Italy	541	2.4	540	2.4	-0.3	0.8
Latvia	545	2.3	544	2.2	-0.5	2.0
Lithuania	543	2.6	544	2.5	0.2	1.3
Macedonia, Rep. of	442	4.6	442	4.8	0.9	1.1
Moldova, Rep. of	492	4.0	492	4.2	0.0	1.0
Morocco	350	9.6	346	10.0	-3.3	1.5
Netherlands	554	2.5	554	2.7	0.1	1.2
New Zealand	529	3.6	529	3.8	0.2	1.4
Norway	499	2.9	500	2.8	0.7	1.1
Romania	512	4.6	512	4.6	-0.1	0.8
Russian Federation	528	4.4	528	4.2	0.0	1.0
Scotland	528	3.6	528	3.5	0.2	1.0
Singapore	528	5.2	528	5.2	0.1	0.6
Slovak Republic	518	2.8	518	2.8	0.2	1.2
Slovenia	502	2.0	502	1.9	0.3	1.3
Sweden	561	2.2	561	2.3	0.1	1.2
United States	542	3.8	542	3.8	0.3	0.9
International average	517	3.4	517	3.4	0.0	1.1
Ontario, Canada	548	3.3	548	3.3	-0.1	1.3
Québec, Canada	537	3.0	538	2.8	0.5	1.1

Table 2 presents average reading achievement and the respective standard errors in PIRLS 2001 and PIRLS 2006 (Mullis et al., 2007). This table also displays the average difference for each participant between the two cycles and its standard error, computed without reference to linking error. Reliance on these traditional standard error estimates in PIRLS 2006 led to 14 countries showing statistically significant changes in reading achievement between 2001 and 2006.

Table 2: Trends in reading achievement

Country	PIRLS 2001 average scale score		PIRLS 2006 average scale score		Difference between PIRLS 2001 and 2006 scores		
	Average achievement	*SE*	Average achievement	*SE*	Difference	*SE* without linking error	*SE* including linking error
Russian Federation [2a]	528	4.4	565	3.4	37	5.6*	5.6*
Hong Kong SAR	528	3.1	564	2.4	36	3.9*	4.0*
Singapore	528	5.2	558	2.9	30	5.9*	5.9*
Slovenia	502	2.0	522	2.1	20	2.9*	3.2*
Slovak Republic	518	2.8	531	2.8	13	4.0*	4.1*
Italy	541	2.4	551	2.9	11	3.8*	3.8*
Germany	539	1.9	548	2.2	9	2.9*	3.0*
Moldova, Rep. of	492	4.0	500	3.0	8	5.0	5.1
Hungary	543	2.2	551	3.0	8	3.7*	3.9*
Iran, Islamic Rep. of	414	4.2	421	3.1	7	5.2	5.4
Canada, Ontario [2a]	548	3.3	554	2.8	6	4.4	4.5
Israel [2b]	509	2.8	512	3.3	4	4.4	4.5
New Zealand	529	3.6	532	2.0	3	4.1	4.3
Macedonia, Rep. of	442	4.6	442	4.1	1	6.2	6.3
Scotland†	528	3.6	527	2.8	-1	4.6	4.7
Norway ‡	499	2.9	498	2.6	-1	3.9	4.0
Iceland	512	1.2	511	1.3	-2	1.8	2.1
United States† [2a]	542	3.8	540	3.5	-2	5.2	5.2
Bulgaria [2a]	550	3.8	547	4.4	-3	5.8	5.9
France	525	2.4	522	2.1	-4	3.1	3.3
Latvia	545	2.3	541	2.3	-4	3.3	3.8
Canada, Québec	537	3.0	533	2.8	-4	4.1	4.2
Lithuania	543	2.6	537	1.6	-6	3.1*	3.3
Netherlands†	554	2.5	547	1.5	-7	2.9*	3.2*
Sweden	561	2.2	549	2.3	-12	3.2*	3.4*
England	553	3.4	539	2.6	-13	4.3*	4.5*
Romania	512	4.6	489	5.0	-22	6.8*	6.8*
Morocco	350	9.6	323	5.9	-27	11.3*	11.4*

Notes:

* $p < 0.05$.

† Met guidelines for sample participation rates only after replacement schools were included.

‡ Nearly satisfied guidelines for sample participation rates after replacement schools were included.

[2a] National defined population covers less than 95% of national desired population.

[2b] National defined population covers less than 80% of national desired population.

Trend note: The primary education systems of the Russian Federation and Slovenia underwent structural changes. Data for Canada, Ontario include public schools only.

The standard error of the difference without including linking error is computed as $SE = \sqrt{SE_1^2 + SE_2^2}$, where SE_1 is the standard error from PIRLS 2001 and SE_2 is the standard error from PIRLS 2006. To include the linking error in the standard error of the difference, the estimate of the linking error for each participant was combined with the existing standard error of the difference as $SE = \sqrt{SE_1^2 + SE_2^2 + SE_L^2}$, where SE_1 and SE_2 are the standard errors from PIRLS 2001 and PIRLS 2006, respectively, and SE_L is the standard error of the link.

Results reflecting the revised standard error estimates are presented in the last column of Table 2. Comparing the published standard errors of the difference between the 2001 and 2006 scale scores to the standard errors that include the linking error demonstrates that the linking error had very little impact on the statistical significance of the PIRLS 2006 trend estimates. In most countries, the standard error that included the linking error increased by less than 0.1 of a point. Due to these small changes, the statistical significance of the trend estimates remained the same for all but one country—Lithuania. When the linking error was set aside, the difference between PIRLS 2001 and PIRLS 2006 average scale scores for Lithuania was -6.35 with a standard error of 3.1, which was a statistically significant decrease in reading achievement between 2001 and 2006. However, when the linking error was included, the standard error became 3.3, making the change in reading achievement not statistically significantly different from zero.

Despite this one change in significance, most participants' trend estimates were not greatly affected by the inclusion of the linking error. For example, Latvia had the highest linking error estimate (2.0). However, adding the linking error to the traditional standard error estimate ($\sqrt{2.3^2 + 2.3^2} = 3.25$) only increased the error by 0.6 points ($\sqrt{2.3^2 + 2.3^2 + 2.0^2} = 3.81$) and did not affect the statistical significance of the trend estimate (i.e., the difference remained nonsignificant).

CONCLUSION AND IMPLICATIONS

As is the case with other large-scale assessments, TIMSS and PIRLS measure trends in student achievement by administering assessments to national student samples every four or five years. About 60% of each assessment is reassessed from previous cycles, and the rest is newly developed. Linking the results of successive assessments to a common achievement scale is a statistical estimation process that necessarily involves some degree of estimation error.

This paper described a procedure for estimating the error in the TIMSS and PIRLS linking process. It also described an application of that procedure to the linking of the PIRLS 2006 data to the PIRLS achievement scale originally established by PIRLS 2001. Because new assessment items are introduced with each assessment cycle as replacements for released items, each cycle requires an item calibration to determine the values of the item parameters necessary to estimate the distribution of student achievement. The fact that item parameters are not constant from cycle to cycle introduces some uncertainty, or linking error, into the trend measurement process.

To minimize linking error, TIMSS and PIRLS use a concurrent calibration process involving items and data from the previous as well as the current assessment and including items common to both assessment cycles. As a result of this process, item parameters are not only determined for all new items but are also reestimated for items common to both the current and previous assessments. The reestimation allows the parameter values of the common items to gradually evolve from assessment cycle to cycle, while providing the best possible fit to the assessment data.

Because the concurrent calibration approach to item parameter estimation uses all available data from both the current and previous assessment cycles, this approach provides the best estimate of the item parameter values. As an additional advantage, it also provides the data to estimate the linking error that results from changes in item parameters from cycle to cycle. The concurrent calibration process results in two sets of item parameter estimates for the items in the "previous" assessment—one set from the calibration conducted for the previous cycle and a second, more recent set from the new concurrent calibration. The differences between the two sets of item parameters can be used to estimate linking error.

For the PIRLS example in this paper, the PIRLS 2001 items had item parameter estimates based on the original 2001 item calibration and a second set based on the 2006 concurrent calibration. Applying the two sets of item parameters to the same data (the PIRLS 2001 data in this paper) provided a measure of the effect on achievement estimation of using one set of parameters in place of the other. The difference between the achievement estimates is a measure of the linking error due to the change in item parameters.

In contrast to the PISA approach to linking error, which is based on the variance of the linking-item difficulty parameters, the TIMSS and PIRLS approach considers linking error in terms of differences in student achievement distributions due to item parameter changes. As such, the latter approach more closely addresses the consequences for student achievement of evolving assessment item pools from one assessment cycle to the next. An advantage of the focus on changes in student achievement distributions is that linking error can be estimated and reported country by country rather than as a single global estimate, as in PISA.

This study indicates that trend estimation in PIRLS 2006 was not greatly affected by including linking error in the computation of standard errors. Most likely, this outcome is a result of the relatively large number of trend items in the PIRLS 2006 assessment design, which reassessed approximately 50% of the assessment from PIRLS 2001 (Martin et al., 2007). The PIRLS linking error, based on 49 common items, averaged 1.1 across the countries, which is relatively close to PISA's estimate of 1.4 for mathematics literacy based on 48 items. These linking error estimates are considerably less than the PISA estimates of 3.1 to 5.3 for reading and science literacy, based on no more than 28 common items.

The findings of this study, and the comparison with PISA results, provide support for the idea that the key to reliable trend measurement lies in having a sufficiently large number of common items in adjacent assessments. Including a component for linking error in tests of achievement differences between two PIRLS assessment cycles with 49 items in common made very little difference to the overall outcome. Nonetheless, to ensure that linking error is further reduced in PIRLS, the number of items common to PIRLS 2006 and PIRLS 2011 has been increased to 75. As indicated earlier, TIMSS also has large numbers of items common to the 2007 and 2011 assessments: 126 and 125 for eighth-grade mathematics and science, respectively, and 103 and 100 for fourth-grade mathematics and science, respectively.

References

Donoghue, J. R., & Mazzeo, J. (1992, April). *Comparing IRT-based equating procedures for trend measurement in a complex test design*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Foy, P., Galia, J., & Li, I. (2007). Scaling the PIRLS 2006 reading assessment data. In M. O. Martin, I. V. S. Mullis, & A. M. Kennedy (Eds.), *PIRLS 2006 technical report* (pp. 149–172). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Johnson, E. (2005). *Trend linking error in PIRLS and TIMSS*. Unpublished manuscript.

Kolen, M., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Lee, W-C., Song, M-Y., & Kim, J-P. (2004, April). *An investigation of procedures for obtaining a common IRT scale*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (Eds.). (2003). *PIRLS 2001 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (Eds.). (2007). *PIRLS 2006 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, *8*(3), 323–335.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muraki, E., Hombo, C. M., & Lee, Y-W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement*, *24*, 325–337.

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

Organisation for Economic Co-operation and Development (OECD). (2009). *PISA 2006 technical report*. Paris, France: OECD Publishing.

Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed.). New York, NY: Springer.