Linking the 2011 National Assessment of Educational Progress (NAEP) in Reading to the 2011 Progress in International Reading Literacy Study (PIRLS)

Gary W. Phillips
American Institutes for Research

February 2014
Commissioned by the NAEP Validity Studies Panel (NVS)

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

This report was prepared for the National Center for Education Statistics under Contract No. ED-IES-13-C-0050 with the American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

1150_02/14

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Peter Behuniak
University of Connecticut

George W. Bohrnstedt
American Institutes for Research

James R. Chromy
Research Triangle Institute

Phil Daro
Strategic Education Research Partnership (SERP) Institute

Lizanne DeStefano
University of Illinois

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Larry Hedges
Northwestern University

Gerunda Hughes
Howard University

Robert Linn
University of Colorado Boulder

Ina V.S. Mullis
Boston College

Scott Norton
Council of Chief State School Officers

Gary Phillips
American Institutes for Research

Lorrie Shepard
University of Colorado Boulder

David Thissen
University of North Carolina, Chapel Hill

Karen Wixson
University of North Carolina, Greensboro

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Janis Brown
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
2800 Campus Drive, Suite 200
San Mateo, CA 94403
Phone: 650/ 843-8100
Fax: 650/ 843-8200

Executive Summary

This paper describes a statistical linking between the 2011 National Assessment of Educational Progress (NAEP) in Grade 4 reading and the 2011 Progress in International Reading Literacy Study (PIRLS) in Grade 4 reading. The primary purpose of the linking study is to obtain a statistical comparison between NAEP (a national assessment) and PIRLS (an international assessment). By expressing both assessments in the same metric, the linking study can provide international benchmarks for the NAEP Grade 4 reading achievement levels. At each level, the linking shows that the NAEP Grade 4 reading achievement levels are higher than the PIRLS international benchmarks. This finding provides one piece of validity evidence that NAEP results are internationally competitive.

CONTENTS

List of Tables

List of Figures

Introduction

This paper shows how the 2011 National Assessment of Educational Progress (NAEP) reading assessment can be linked to the Progress in International Reading Literacy Study (PIRLS) by placing NAEP and PIRLS on the same scale. Conceptually, linking two assessments simply means the two are connected in such a way that there is a cross-walk between them (e.g., a cross-walk between NAEP Grade 4 reading and PIRLS), thereby allowing one to see where a score point on one of the assessments would fall on the scale of the other assessment.

Linking is a statistical procedure that allows one to express the results of one test (e.g., the 2011 NAEP reading assessment) in terms of the metric of another (e.g., the 2011 PIRLS). The comparisons are valid to the extent that the test content is similar between the two tests. (Evaluating the similarity is usually a judgment call and not a statistical decision.) Using an analogy from the physical sciences, linking is similar to expressing Celsius in terms of Fahrenheit. The cross-walk is the equation $F^o = 32 + 1.8\left(C^o\right)$. In this equation, 32 and 1.8 are the intercept and slope, respectively, of a straight line. Both the intercept and slope are known without error. The cross-walk between the 2011 NAEP Grade 4 reading assessment and the 2011 PIRLS is similar, but because the intercept and slope must be empirically estimated from sample data, they are subject to error, unlike the cross-walk between temperature metrics. The determination of this cross-walk, and error, are the primary outcomes of statistical linking studies.

The paper is organized as follows. First, we provide a brief review of related background literature in which various linking methods are described. A more extended description of statistical moderation (which is the linking method used in this paper) follows. We then present the results of the linking, along with information on conducting the calculations using NAEP plausible values and estimating the linking parameters (along with their standard errors). Finally, we compare the NAEP Grade 4 reading achievement levels with the international benchmarks in PIRLS. We do not predict state PIRLS results from state NAEP reading results because it was not possible to estimate the correlation between state NAEP and state PIRLS.

Background

Linking Methods

Mislevy (1992) and Linn (1993) have described many of the conceptual and statistical issues associated with linking assessments. They have designated four types of statistical linking: *equating*, *calibration*, *projection*, and *statistical moderation*. A brief explanation of the differences is provided here.

In *equating,* both tests, *X* and *Y*, have been designed and developed to be equally reliable, and each measures the same content. Equating is most often used when the goal is to relate two alternate forms of the same test, such as alternate forms of the ACT or the SAT. In equating, the distributions of tests *X* and *Y* are aligned or matched up directly. The matching can be done with equi-percentile equating or linear equating, and the distributions can either be observed score distributions or estimates of the true score distributions. When the equating assumptions (same content and equal reliability) are met:

- The linking function should be the same for *X* expressed in terms of *Y* and for *Y* expressed in terms of *X*.
- The linking function should be the same for different subgroups, across contexts and time.

In *calibration* (for example, with the use of item-response theory), two tests are assumed to measure the same content, but they are not equally reliable. For example, one test, *X,* might be a long test, whereas the other test, *Y,* might be short. The two versions of the test are not equated, but they are indirectly comparable because they have been calibrated to a common scale, θ. This type of linking is done across years in NAEP, PIRLS, the Trends in International Mathematics and Science Study (TIMSS), the Program for International Student Assessment (PISA), most state criterion-referenced tests, as well as most nationally standardized norm-referenced tests. Calibration procedures provide unbiased estimates for individual students and means, but additional statistical machinery is needed to accurately estimate group characteristics such as the variance or the percentages at and above achievement levels. When the same content assumption is met:

- The linking function between *X* and θ (e.g., the test characteristic curve) is different from the linking function between *Y* and θ.
- Both *X* and *Y* can be used to get unbiased estimates of θ for individual students (although the conditional standard error of measurement is different).
- However, the observed score distributions of *X* for groups do not match the observed score distributions for *Y* unless *X* and *Y* are equally reliable.

In *projection,* a regression equation uses the correlation between the two tests to predict the scores on one test, *Y,* from those of another test, *X*. There is no assumption that the two tests measure the same content or that they are equally reliable. However, there is an assumption that the tests are highly correlated. With

projection, there is no longer a symmetric relationship between one test and the other. The conversion table for predicting the first test from the second is different from the table predicting the second test from the first. Following are some of the features of projection:

- The linking function for *X* expressed in terms of *Y* (e.g., regression equation) will be different from the linking function for *Y* expressed in terms of *X*.
- The linking function will likely be different for different subgroups and across contexts and time.

In *statistical moderation*, the scores on the first test, *X,* are adjusted to have the same distributional characteristics as the scores on the second test, *Y*. In this case, assume *X* is linked to *Y*. This is typically done by matching the means and standard deviations of *X* and *Y*, or matching their percentile ranks. The usual requirement for statistical moderation is that both *X* and *Y* have been administered to comparable populations of students (e.g., the student populations taking the two tests are randomly equivalent). When statistical moderation is used:

- The linking function for *X* expressed in terms of *Y* (e.g., a z-score equivalency) will be different from the linking function for *Y* expressed in terms of *X*.
- The linking function will likely be different for different subgroups, across contexts and time.
- The degree of the relationship between *X* and *Y* is typically unknown.

Holland (2007) provided a more recent categorization of linking methods. In this categorization, linking is divided into *predicting*, *scale aligning* (or scaling), and *equating*.

In *predicting* observed scores, one is predicting the observed test score *Y* from the observed test score *X* (which might be several tests as well as demographic information) in population *P* through the equation $y = \mathrm{E}(Y \mid X = x, P)$. In addition to predicting individual observed scores, one also can project the distribution of scores from the scale of *X* to the scale of *Y* in population *P* through the conditional cumulative distribution function $\Pr\left(\mathbf{Y} \le y \mid \mathbf{X} = x, \mathbf{P}\right)$. Once the projection equation has been determined for *P* (the linking sample), where both *Y* and *X* have been administered, it can be applied in population *Q,* where *X* has been administered but *Y* has not been administered. Pashley and Phillips (1993) used this methodology to project the distribution of scores from the International Assessment of Educational Progress (IAEP) to the NAEP scale.

In *scale alignment*, the scores on *X* are rescaled to be in the metric of *Y*. *Vertical scaling* and *calibration* are examples of scale alignment.

In *equating,* the methodology may be the same as linking, but the procedure must meet the following additional requirements:

- *Equal constructs*—the tests must measure the same construct.
- *Equal reliability*—the tests must be equally reliable.

- *Symmetry*—the equating function must be symmetrical (the function used to equate Y to X must be the inverse of the function for equating X to Y).
- *Equity*—it should not matter which test is administered to the examinee.
- *Population invariance*—the equating function between X and Y should be the same in all populations.

Linking NAEP to International Assessments

Several major attempts have been made to link NAEP statistically to international assessments.

The first attempt involved linking the 1991 IAEP to the 1992 NAEP in mathematics (Pashley & Phillips, 1993). The IAEP was first conducted in February 1988 in five countries (Ireland, Korea, Spain, the United Kingdom, and the United States) and four provinces in Canada using representative samples of 13-year-old students assessed in mathematics and science (LaPointe, Mead, & Phillips, 1989). The IAEP was expanded and repeated again in 1991 in 20 countries in which representative samples of 9- and 13-year-old students were assessed in mathematics and science (LaPointe, Meade, & Askew, 1992). Pashley and Phillips (1993) conducted the IAEP-NAEP linking study in mathematics using *projection* methodology. To establish the link between IAEP and NAEP mathematics, a nationally representative linking sample of 1,609 students was administered both IAEP and NAEP in 1992. The linking study used samples of eighth-grade students who took NAEP and 13-year-old students who took IAEP. (NAEP was based on grade whereas IAEP was based on age.) The direction of the link was to predict NAEP performance from IAEP results in other countries. The purpose of the study was to estimate how other countries stacked up against the NAEP achievement levels in mathematics. The IAEP-NAEP linkage was conducted within the context of the policy environment at the time. The nation's governors, along with the President, had held the National Education Summit and adopted six broad national goals. The fourth goal was that, by the year 2000, "U.S. students would be the first in the world in science and mathematics achievement." The IAEP-NAEP linking study was the first effort to directly address the need for a common metric and common standard in international comparisons (i.e., predicting how other countries would do on NAEP based on their performance on IAEP). Once the predicted NAEP scores were obtained, the NAEP achievement levels were used to report different countries' performance. The IAEP was not repeated; however, it had many design features (such as linking studies) that were incorporated into subsequent international assessments such as PIRLS.

A second attempt to link NAEP to an international study was done by Beaton and Gonzales (1993). They used *statistical moderation* to link the 1991 IAEP to the 1990 NAEP scale in mathematics. The results of the Beaton and Gonzales study were similar to the Pashley and Phillips (1993) study only for countries with performance similar to the U.S. average.

The third study used *statistical moderation* to link the Grade 4 and Grade 8 1996 NAEP to the Grade 4 and Grade 8 1995 TIMSS in mathematics and science (Johnson &

Siengondorf, 1998). Based on the validation analyses (in two states that administered both NAEP and TIMSS), the NAEP-TIMSS link appeared to work at Grade 8 but not at Grade 4.[1]

The fourth study (Johnson, Cohen, Chen, Jiang, & Zhang, 2005) used *projection* methods (similar to Pashley and Phillips, 1993) for Grade 8 mathematics and science to link NAEP to TIMSS. The TIMSS assessment in mathematics and science was conducted in 1999, and the NAEP assessment in mathematics and science was conducted in 2000. In addition to projection methods, the study also used *statistical moderation* as a secondary method of linking. Based on a validation study in which 12 states took both NAEP and TIMSS, the general finding was that, for the U.S. national linking sample, the projection method did not work. However, the statistical moderation method (which used the national samples of both NAEP and TIMSS instead of the linking sample) did perform well in the validation study.

One important caveat with these analyses is that the standard errors and the validation analyses are based on data collected only within the United States. In the United States, students took both NAEP and TIMSS. In all other countries, however, students only took TIMSS. Whether the linking parameters are stable in other countries is an empirical question that the study by Johnson and colleagues (2005) could not answer. In fact, no international linking study has been designed to answer this question. There is no guarantee that linking parameters estimated from one nation (e.g., the United States) will be the same as those in other nations.

[1] The link worked at Grade 8 based on the validation sample. The predicted TIMSS results for Minnesota (the only state that administered the eighth-grade TIMSS) were comparable to the actual TIMSS results. The link did not work at Grade 4. The predicted TIMSS results for the two states that administered the fourth-grade TIMSS (Colorado and Minnesota) were considerably higher than the actual TIMSS results. The study was not able to determine why this result occurred in the Grade 4 link.

Method

Statistical Moderation

In the study reported here, *X* = NAEP was linked to *Y* = PIRLS using statistical moderation and resulting in NAEP reading scores expressed on the PIRLS scale—referred to as $l_Y(x_i)$. This means that the estimated PIRLS scores $l_Y(x_i)$ are actually NAEP reading scores adjusted to have the same mean and standard deviation as PIRLS. That is what it means in *statistical moderation* to say "NAEP is linked to PIRLS."

The linear PIRLS-equivalent $l_Y(x_i)$ associated with a NAEP reading score point x_i is

$$l_Y(x_i) = \left(\overline{y} - \frac{\hat{\sigma}_y}{\hat{\sigma}_x}\overline{x}\right) + \left(\frac{\hat{\sigma}_y}{\hat{\sigma}_x}\right)x_i \tag{1}$$

$$\hat{A} = \overline{y} - \frac{\hat{\sigma}_y}{\hat{\sigma}_x}\overline{x}$$
$$\hat{B} = \frac{\hat{\sigma}_y}{\hat{\sigma}_x} \tag{2}$$

In equations (1) and(2),

- $\hat{A}$ is an estimate of the intercept of a straight line, and $\hat{B}$ is an estimate of the slope.
- $\overline{x}$ and $\overline{y}$ are the national public school means of the U.S. 2011 NAEP reading assessment and 2011 U.S. PIRLS results.
- $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are the public school standard deviations for NAEP reading and PIRLS, respectively.
- $l_Y(x_i)$ is the observed score on *X* (NAEP) converted to the scale of *Y* (PIRLS).

Linking Error Variance

With statistical moderation, $l_Y(x_i)$ is a linear transformation of x_i. Therefore, the error variance in $l_Y(x_i)$ is

$$\hat{\sigma}^2_{l_Y(x_i)} = \hat{B}^2\hat{\sigma}^2_{x_i} + Var\left(\hat{A}\right) + 2(x_i)Cov\left(\hat{A},\hat{B}\right) + (x_i)^2 Var\left(\hat{B}\right). \tag{3}$$

According to Johnson et al. (2005), the error variances of the parameters of the linear transformation, $Var\left(\hat{A}\right), Cov\left(\hat{A}, \hat{B}\right),$ and $Var\left(\hat{B}\right)$ can be approximated by Taylor-series linearization (Wolter, 1985).

$$
\begin{aligned}
Var\left(\hat{A}\right) &= \hat{B}^2\hat{\sigma}_{\bar{x}}^2 + \hat{\sigma}_{\bar{y}}^2 + \bar{x}^2\hat{B}^2\left[\frac{Var\left(\sigma_Y\right)}{\hat{\sigma}_Y^2} + \frac{Var\left(\sigma_X\right)}{\hat{\sigma}_X^2}\right] \\
Cov\left(\hat{A}, \hat{B}\right) &= -\bar{x}\hat{B}^2\left[\frac{Var\left(\sigma_Y\right)}{\hat{\sigma}_Y^2} + \frac{Var\left(\sigma_X\right)}{\hat{\sigma}_X^2}\right] \\
Var\left(\hat{B}\right) &= \hat{B}^2\left[\frac{Var\left(\sigma_Y\right)}{\hat{\sigma}_Y^2} + \frac{Var\left(\sigma_X\right)}{\hat{\sigma}_X^2}\right].
\end{aligned}
\tag{4}
$$

Results

Parameter Estimates of the Mean and Standard Deviation

In this study, only public school students were included in the analysis of plausible values for both NAEP reading and PIRLS. In both NAEP reading and PIRLS, five plausible values are used to represent a student's posterior distribution. The parameter being estimated is labeled as *P,* the number of plausible values as *N,* and the estimates of *P* as p_n, for $n = 1, 2, ..N$. The average of the statistics is $\overline{p}$, where $\overline{p} = \sum_{n=1}^{N} \frac{p_n}{N}$. Table 1 below shows the calculations for the parameter estimates of the means and standard deviations.

Table 1. Estimating the Mean and Standard Deviation in 2011 U.S. National Samples (Public Schools) for Grade 4 Reading

	Plausible Value 1	Plausible Value 2	Plausible Value 3	Plausible Value 4	Plausible Value 5	Mean Plausible Value ($\overline{p}$)
NAEP reading mean	220.03	220.04	219.99	220.07	220.00	220.026
PIRLS reading mean	556.73	556.05	556.00	556.36	556.73	556.375
NAEP reading standard deviation	36.01	36.01	36.11	36.05	36.07	36.050
PIRLS reading standard deviation	73.34	73.72	73.12	73.62	73.36	73.431

Error Variance (Sampling) of the Mean and Standard Deviation

The error variances for the parameter estimates each have two components—error variance due to sampling (S) and error variance due to measurement (M). The sampling error in the estimates of the means and standard deviations was obtained by using a jackknife error variance approach for complex samples. The jackknife procedure was carried out for each plausible value and then averaged across all five plausible values. In the jackknife procedure, one primary sampling unit (PSU) is excluded, the sampling weights are redistributed across the other units within the stratum in which the PSU was excluded, the mean and standard deviation are calculated on the remaining PSUs, and the process is repeated until all PSUs have been excluded. After the jackknife procedure is carried out on each plausible value, the average across plausible values is $S = \sum_{n=1}^{N} \frac{S_n}{N}$.

This process results in the variance estimates reported in Table 2, which are estimates of error variance due to sampling for the means and standard deviations.

Table 2. Sampling Error Variance of the Mean and Standard Deviation $\left(S_\mu, S_\sigma\right)$ for Grade 4 Reading

Variance of NAEP mean 2011 reading from jackknife	0.095
Variance of PIRLS mean 2011 reading from jackknife	2.216
Variance of NAEP standard deviation 2011 reading from jackknife	0.022
Variance of PIRLS standard deviation 2011 reading from jackknife	0.834

Error Variance (Measurement) of the Mean and Standard Deviation

The error variance due to measurement is estimated by the variance between plausible values. This is estimated by $M = \frac{1+\left(1/N\right)}{N-1}\sum_{n=1}^{N}\left(p_n - \overline{p}\right)^2$. The error variance due to measurement is shown in Table 3.

Table 3. Measurement Error Variance of the Mean and Standard Deviation $\left(M_\mu, M_\sigma\right)$ for Grade 4 Reading

Variance of NAEP mean 2011 reading from plausible values	0.001
Variance of PIRLS mean 2011 reading from plausible values	0.147
Variance of NAEP standard deviation 2011 reading from plausible values	0.002
Variance of PIRLS standard deviation 2011 reading from plausible values	0.069

Error Variance (Total) of the Mean and Standard Deviation

The total error variance is $T = S + M$ and is displayed in Table 4.

Table 4. Total Error Variance of the Mean and Standard Deviation $\left(T_\mu, T_\sigma\right)$ for Grade 4 Reading

Variance of NAEP mean 2011 reading	0.096
Variance of PIRLS mean 2011 reading	2.363
Variance of NAEP standard deviation 2011 reading	0.024
Variance of PIRLS standard deviation 2011 reading	0.903

Linking Parameters

The linking parameters are calculated for each plausible value, using equation (2). The linking parameter estimates are then averaged over the five plausible values as reported in Table 5.

Table 5. Estimating the Linking Parameters A and B in the U.S. National Sample (Public Schools) for Grade 4 Reading

	Plausible Value 1	Plausible Value 2	Plausible Value 3	Plausible Value 4	Plausible Value 5	Mean Plausible Value ($\overline{p}$)
$\hat{A}$	108.609	105.672	110.516	106.877	109.324	108.201
$\hat{B}$	2.037	2.047	2.025	2.042	2.034	2.037

Error Variance (Sampling) of the Linking Parameters A and B

The error variances of the linking parameter estimates $\hat{A}$ and $\hat{B}$ are found using equation (4). These error variances also have two components—one due to sampling and one due to measurement error. The error variances due to sampling are reported in Table 6.

Table 6. Sampling Error Variance in NAEP–PIRLS Linking Parameters for Grade 4 Reading $\left(S_A, S_B, S_{AB}\right)$

Error variance in A, $\left(\hat{\sigma}^2_{A(s)}\right)$	37.077
Two times the covariance between A and B, $2\left(\hat{\sigma}_{AB(s)}\right)$	-0.157
Error variance in B, $\left(\hat{\sigma}_{B(s)}\right)$	0.001

Error Variance (Measurement) of the Linking Parameters A and B

The error variances in the linking parameters due to measurement are displayed in Table 7.

Table 7. Measurement Error Variance in NAEP–PIRLS Linking Parameters for Grade 4 Reading $\left(M_A, M_B, M_{AB}\right)$

Error variance in A, $\left(\hat{\sigma}^2_{A(m)}\right)$	3.037
Two times the covariance between A and B, $2\left(\hat{\sigma}_{AB(m)}\right)$	-0.013
Error variance in B, $\left(\hat{\sigma}_B\right)$	0.000

Error Variance (Total) of the Linking Parameters A and B

The sum of the error variances due to sampling and the error variances due to measurement are reported in Table 8.

Table 8. Total Error Variance in NAEP–PIRLS Linking Parameters for Grade 4 Reading (T_A, T_B, T_{AB})

Error variance in A, $\left(\hat{\sigma}_A^2\right)$	40.114
Two times the covariance between A and B, $2\left(\hat{\sigma}_{AB}\right)$	-0.170
Error variance in B, $\left(\hat{\sigma}_B\right)$	0.001

Linking Error Variance of the Projected NAEP Reading Achievement Levels

The linking error variance of the projected NAEP reading achievement levels on the PIRLS scale is found in equation (3). The linking error variance also has two components—one due to sampling and one due to measurement error. The sampling components are included in Table 9 and the measurement components are displayed in Table 10.

Table 9. Error Variance in Linking Due to Sampling for NAEP Reading Achievement Levels Projected Onto the PIRLS Grade 4 Reading Scale

$\hat{\sigma}_{PIRLS_{adv}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{adv}} + \hat{\sigma}^2_{A(s)} + 2\left(NAEP_{adv}\right)\hat{\sigma}_{AB(s)} + \left(NAEP_{adv}\right)^2\hat{\sigma}^2_{B(s)}$	4.245
$\hat{\sigma}^2_{PIRLS_{prof}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{prof}} + \hat{\sigma}^2_{A(s)} + 2\left(NAEP_{prof}\right)\hat{\sigma}_{AB(s)} + \left(NAEP_{prof}\right)^2\hat{\sigma}^2_{B(s)}$	2.839
$\hat{\sigma}^2_{PIRLS_{basic}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{basic}} + \hat{\sigma}^2_{A(s)} + 2\left(NAEP_{basic}\right)\hat{\sigma}_{AB(s)} + \left(NAEP_{basic}\right)^2\hat{\sigma}^2_{B(s)}$	2.712

Table 10. Error Variance in Linking due to Measurement for NAEP Reading Achievement Levels Projected Onto the PIRLS Grade 4 Reading Scale

$\hat{\sigma}_{PIRLS_{adv}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{adv}} + \hat{\sigma}^2_{A(m)} + 2\left(NAEP_{adv}\right)\hat{\sigma}_{AB(m)} + \left(NAEP_{adv}\right)^2\hat{\sigma}^2_{B(m)}$	0.288
$\hat{\sigma}^2_{PIRLS_{prof}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{prof}} + \hat{\sigma}^2_{A(m)} + 2\left(NAEP_{prof}\right)\hat{\sigma}_{AB(m)} + \left(NAEP_{prof}\right)^2\hat{\sigma}^2_{B(m)}$	0.170
$\hat{\sigma}^2_{PIRLS_{basic}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{basic}} + \hat{\sigma}^2_{A(m)} + 2\left(NAEP_{basic}\right)\hat{\sigma}_{AB(m)} + \left(NAEP_{basic}\right)^2\hat{\sigma}^2_{B(m)}$	0.160

The sum of the linking error variance due to sampling in Table 9 and the linking error variance due to measurement in Table 10 yields the total linking error variances in the projected achievement levels on the PIRLS scale reported in Table 11.

Table 11. Total Error Variance in Linking for NAEP Reading Achievement Levels Projected Onto the PIRLS Grade 4 Reading Scale

$\hat{\sigma}_{PIRLS_{adv}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{adv}} + \hat{\sigma}^2_A + 2\left(NAEP_{adv}\right)\hat{\sigma}_{AB} + \left(NAEP_{adv}\right)^2\hat{\sigma}^2_B$	4.536
$\hat{\sigma}^2_{PIRLS_{prof}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{prof}} + \hat{\sigma}^2_A + 2\left(NAEP_{prof}\right)\hat{\sigma}_{AB} + \left(NAEP_{prof}\right)^2\hat{\sigma}^2_B$	3.009
$\hat{\sigma}^2_{PIRLS_{basic}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{basic}} + \hat{\sigma}^2_A + 2\left(NAEP_{basic}\right)\hat{\sigma}_{AB} + \left(NAEP_{basic}\right)^2\hat{\sigma}^2_B$	2.872

One interesting question in linking studies is, "How much of the linking error is due to sampling and how much is due to test unreliability (or measurement error)?" In this study, we can answer that question by comparing the error variances in Table 9 (sampling error in linking) and Table 10 (measurement error in linking) with Table 11 (total error in linking).Table 12 shows the percentages of linking error variance accounted for by sampling and measurement error, respectively, for the NAEP reading achievement levels.

Table 12. Variance Components of Linking Error for NAEP Reading Achievement Levels Projected Onto the PIRLS Grade 4 Reading Scale

	Sampling	Measurement
Advanced	93.6%	6.4%
Proficient	94.3%	5.7%
Basic	94.4%	5.6%

As Table 12 clearly shows, the vast majority of linking error is due to sampling. However, measurement error becomes a larger percentage of the linking error in the tails of the achievement distribution. This is why the measurement error for the advanced achievement level is larger than for the other two achievement levels.

The means and standard deviations summarized from the plausible values analyses above are reported in Table 13, and the resulting estimates of the linking parameters are reported in Table 14.

Table 13. Means and Standard Deviations for National Samples of Grade 4 U.S. Public School Students, 2011 PIRLS and 2011 NAEP Reading Assessment

	2011 NAEP Reading		2011 PIRLS	
	Mean	SD	Mean	SD
Statistic	220	36	556	73
SE	0.10	0.02	2.36	0.90

Note: SD=standard deviation; SE=standard error

Table 14. Slope and Intercept for Estimating 2011 PIRLS Scores From the 2011 NAEP Reading Assessment Using Statistical Moderation With U.S. National Samples

	A	*B*
Parameter	108.20	2.04
SE	6.33	0.03
Covariance	-0.17	

Note: SE=standard error

The NAEP reading achievement levels projected onto the PIRLS scale are summarized in Table 15, and the four PIRLS benchmarks projected onto the NAEP reading scale are summarized in Table 16. The linking parameters for linking PIRLS to NAEP reading are presented in Appendix A.

Table 15. Grade 4 2011 NAEP Reading Achievement Levels Linked to Grade 4 2011 PIRLS

	NAEP Reading Achievement Level	PIRLS Equivalent	Standard Error of PIRLS Equivalent
Advanced	268	654	2.2
Proficient	238	593	1.9
Basic	208	532	1.8

Table 16. Grade 4 2011 PIRLS Reading International Benchmarks Linked to the Grade 4 2011 NAEP Reading Assessment

	PIRLS International Benchmark	NAEP Reading Equivalent	Standard Error of NAEP Reading Equivalent
Advanced	625	254	1.2
High	550	217	1.1
Intermediate	475	180	1.2
Low	400	143	1.5

In each case the NAEP reading achievement levels are higher than the PIRLS international benchmarks. This can be seen more clearly in Figure 1 where the NAEP achievement levels are expressed in the PIRLS metric and in Figure 2 where the PIRLS international benchmarks are expressed in the NAEP metric.

Figure 1. Comparing NAEP Reading Achievement Levels With PIRLS International Benchmarks

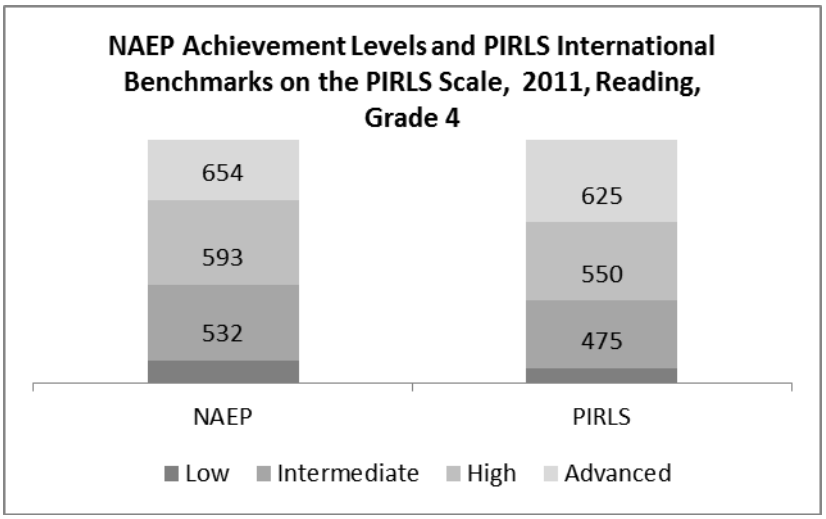
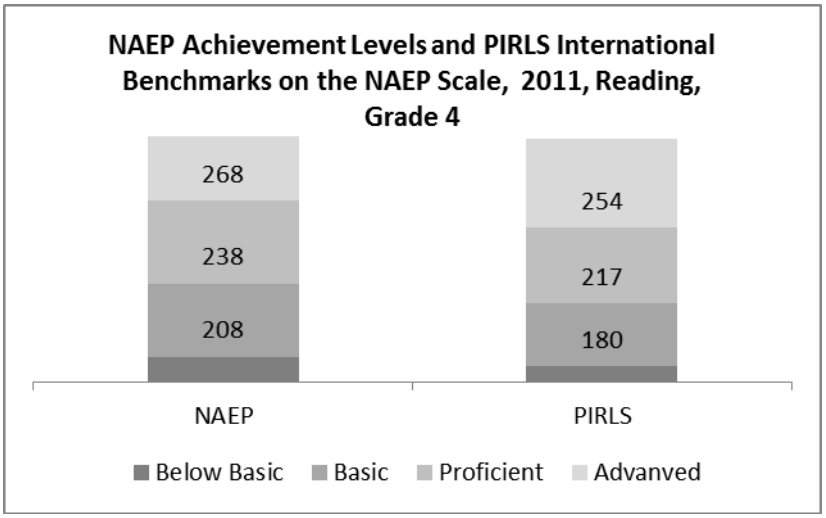


Figure 2. Comparing PIRLS International Benchmarks With NAEP Reading Achievement Levels



The fact that NAEP reading achievement levels are higher than similar PIRLS international benchmarks may help explain why NAEP has historically reported lower rates of reading proficiency for the United States, whereas PIRLS has historically reported higher levels of reading proficiency. For example, in 2011, NAEP reported that 34 percent of fourth graders were reading at the proficient level, while PIRLS reported that 56 percent were reading at the high international benchmark.

Validation

State Samples

The linking between NAEP reading and PIRLS was conducted in the 2011 U.S. national public school samples for both NAEP and PIRLS. One measure of validity (population invariance) would be to apply the linking parameters obtained in the national samples to a state sample. In 2011, Florida was the only state that administered a statewide assessment in PIRLS. This provided a piece of validity evidence on how well the PIRLS-Equivalent obtained from the NAEP-PIRLS linking could be applied to state results. In general, the estimated PIRLS is not significantly different from the actual PIRLS. For example, the mean difference between the PIRLS-Equivalent and the actual PIRLS mean is not significant (see Table 17). The only significant difference between the PIRLS-Equivalent and the actual PIRLS result is for the percentage of advanced students (see Table 18). Even though there is only a 1 percent difference between the predicted and the actual percentage, the difference is statistically significant because the standard errors are so small.

Table 17. Comparing Means for the State PIRLS-Equivalent With the Actual State PIRLS

Florida	PIRLS-Equivalent State Mean	Standard Error Linking	Actual PIRLS State Mean	Standard Error State PIRLS	Overall Standard Error	*Z*-Test	Significant Difference
Mean	566	2.8	569	2.9	4.0	-0.83	NS

Note: Two-tailed *z*-test, with alpha=.05; NS=not significant

Table 18. Comparing Percentages Above Benchmarks for the State PIRLS-Equivalent With the Actual State PIRLS

Florida	PIRLS-Equivalent of State Percentage	Standard Error PIRLS-Equivalent	Actual PIRLS State Percentage	Standard Error State Percentage	Overall Standard Error	*Z*-Test	Significant Difference
Advanced	18	1.9	22	1.7	2.5	-1.40	NS
High	59	2.4	61	1.7	2.9	-0.59	NS
Intermediate	91	1.2	91	1.1	1.7	0.24	NS
Low	99	0.2	98	0.4	0.5	3.04	Significant

Note: Two-tailed *z*-test, with alpha=.05; NS=not significant

Estimation Procedures for Standard Errors

The standard errors of the linking parameters in the NAEP PIRLS linking study in reading were obtained through Taylor Series linearization. One validity question is how would the standard errors compare if a different estimation procedure were used? To test this, the standard errors of the linking parameters also were estimated using the Jackknife method (see Table 19). A more detailed explication of the procedure is included in Appendix B. Because PIRLS has 75 replicate weights compared with 62 in NAEP, we used the first 62 of the 75 replicate weights from PIRLS. We found the standard errors change slightly but not substantially if you use a different set of 62 replicates out of 75.

Table 19. Slope and Intercept for Estimating 2011 PIRLS Scores From the 2011 NAEP Reading Assessment Using Jackknife Estimation Versus Taylor Series With U.S. National Samples

Jackknife		
	A	***B***
Parameter	108.97	2.04
SE	6.32	0.03
Covariance	-0.02	
Taylor Series		
	A	*B*
Parameter	108.20	2.04
SE	6.33	0.03
Covariance	-0.17	

Note: SE=standard error

Note that the results from the Taylor Series are very close to those from the Jackknife procedure. The small discrepancies between the Taylor Series and the Jackknife are likely due to which 62 replicate weights were selected for PIRLS.

Caveats

There are several important caveats to keep in mind regarding the linking of the 2011 NAEP Grade 4 reading assessment to PIRLS.

First, one should consider the content similarities and differences between NAEP reading and PIRLS. Identical content between the two tests is not a requirement for linking. However, the two tests should measure similar content and the more similar the better. A paper comparing the content of the two assessments reported that "The comparison of the NAEP and PIRLS fourth-grade reading assessments suggests that there is a great deal of overlap in what the two assessments are measuring" (Binkley, 2003, p. 26). That is, the two assessments appear to measure similar content.

A second consideration is the similarities and differences in the demographics of the linking samples taking the NAEP reading assessment and PIRLS. Although both NAEP reading and PIRLS were administered in the United States in 2011, NAEP reading was administered earlier than PIRLS, between January and March, while PIRLS was administered between April and June.

There also were differences in the target population due to differential exclusion rates. The overall exclusion rate for PIRLS was 7.2 percent. Because NAEP allows many accommodations, the NAEP Grade 4 reading exclusion rate was smaller—about 4 percent. The fact that the exclusion rates are different indicates that the two target populations are slightly different. However, in a larger 2011 NAEP-TIMSS linking study conducted by the National Center for Education Statistics (NCES) and American Institutes for Research (AIR), adjustments for exclusion rates did not greatly change the results (National Center for Education Statistics, 2013).

The limited validity information available for the present study comes from Florida, which administered both NAEP reading and PIRLS. The comparisons between the estimated PIRLS results and the actual PIRLS results in Florida show that the linking results did a good job of predicting the state results. However, NCES felt that the amount of validity evidence, being limited to only one state, was not sufficient to evaluate whether it was appropriate to use the linking to predict state PIRLS results for all 50 states. This is in contrast with the NAEP-TIMSS linking study in Grade 8 mathematics and science conducted in 2011, which used data from nine states for validation. In the NAEP-TIMSS study, the linking results were used to predict TIMSS performance in all 50 states.

References

Beaton, A. E., & Gonzales, E. J. (1993). *Comparing the NAEP trial state assessment results with the IAEP international results* (report prepared for the National Academy of Education Panel on the NAEP Trial State Assessment). Stanford, CA: National Academy of Education.

Binkley, M. A. (2003). *Content comparison of the NAEP and PIRLS fourth-grade reading assessments* (Working Paper No. 2003-10). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Holland, P (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales*, New York, NY: Springer.

Johnson, E. G., Cohen, J., Chen, W.-H. Jiang, T., & Zhang, Y. (2005). *2000 NAEP–1999 PIRLS linking report* (Publication No. 2005-01). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Johnson, E. G., & Siengondorf, A. (1998). Linking the National Assessment of Educational Progress and the Third International Mathematics and Science Study: Eighth grade results. (Publication No. NCES 98-500). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

LaPointe, A. E., Mead, N. A., & Askew, J. M. (1992). *Learning mathematics*. Princeton, NJ: Educational Testing Service.

LaPointe, A. E., Mead, N. A., & Phillips, G. W. (1989). *A world of differences: An international assessment of mathematics and science* (Report No. 19-CAEP-01). Princeton, NJ: Educational Testing Service.

Linn, R. L. (1993). Linking results of district assessments. *Applied Measurement in Education, 6*, 83–102.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods and prospects.* Princeton, NJ: Policy Information Center, Educational Testing Service.

National Center for Education Statistics. (2013). *The Nation's Report Card: U.S. states in a global context: Results from the 2011 NAEP-TIMSS linking study* (NCES 2013-460). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Pashley, P. J., & Phillips, G. W. (1993). Toward world class standards: A research study linking national and international assessments. Princeton, NJ: Educational Testing Service.

Wolter, K. (1985). *Introduction to variance estimation*. New York, NY: Springer-Verlag.

Appendix A. Linking 2011 PIRLS to 2011 NAEP Grade 4 Reading

This report has been concerned with using statistical moderation to link the 2011 NAEP reading assessment to 2011 PIRLS, thereby expressing NAEP reading results in the PIRLS metric. The linking parameters, along with their standard errors, are contained in Table 14. However, one also can use the same data to express PIRLS scores in the NAEP metric. The linking parameters for this are shown in Table 20.

Table 20. Slope and Intercept for Estimating 2011 NAEP Grade 4 Reading Scores From 2011 PIRLS, Using Statistical Moderation With U.S. National Samples

	A	*B*
Parameter	-53.12	0.49
Standard Error	3.81	0.01
Covariance	-0.02	

Appendix B. Jackknife Standard Errors of Linking Parameters

Burhan Ogut, *American Institutes for Research*

Let

$$Y_{ij} = A_{ij} + B_{ij}X_{ij}$$

represent the function linking NAEP reading scores (X) to PIRLS scores (Y), where Y_{ij} is the PIRLS scale score for the i^{th} plausible value with j^{th} survey weight. The intercept parameter, A_{ij}, and the slope parameter, B_{ij}, of the linking function for the i^{th} plausible value with j^{th} survey weight is computed as,

$$A_{ij} = \bar{y}_{ij} - \frac{\hat{\sigma}_{yij}}{\hat{\sigma}_{xij}}\bar{x}_{ij}$$

$$B_{ij} = \frac{\hat{\sigma}_{yij}}{\hat{\sigma}_{xij}}$$

where $\bar{y}_{ij}$ is the mean PIRLS scale score for the i^{th} plausible value using j^{th} total survey weight and $\hat{\sigma}_{yij}$ is the standard deviation of that mean. Similarly, x_{ij} is the mean NAEP reading scale score for the i^{th} plausible value using j^{th} total survey weight and $\hat{\sigma}_{xij}$ is the standard deviation of this mean.

The jackknife estimates of the A and B parameters are computed as

$$\hat{A} = \sum_{i=1}^{5} \bar{y}_{i,j=total\ survey\ weight} - \frac{\hat{\sigma}_{yi,j=total\ survey\ weight}}{\hat{\sigma}_{xi,j=total\ survey\ weight}}\bar{x}_{i,j=total\ survey\ weight}$$

$$\hat{B} = \sum_{i=1}^{5} \frac{\hat{\sigma}_{yi,j=total\ survey\ weight}}{\hat{\sigma}_{xi,j=total\ survey\ weight}}$$

where $\bar{y}_{i,j=total\ survey\ weight}$ is the mean PIRLS scale score for the i^{th} plausible value using the total survey weight (i.e., TOTWGT) and $\hat{\sigma}_{yi,j=total\ survey\ weight}$ is the standard deviation of this mean. Similarly, $x_{i,j=total\ survey\ weight}$ is the mean NAEP reading scale score for the i^{th} plausible value using the total survey weight (i.e., ORIGWT) and $\hat{\sigma}_{xi,j=total\ survey\ weight}$ is the standard deviation of this mean.

The error variance component of the jackknife estimates of the A and B parameters has two subcomponents: measurement and sampling.

Measurement error variance is the variance between the estimates of the A and B parameters with five plausible values using the total survey weight. The measurement error variances of the A and B parameters are computed as

$$M_A = Var(\hat{A}) = \frac{1}{5-1}\sum_{i=1}^{5}(\hat{A}_i - \bar{A})^2$$

$$M_B = Var(\hat{B}) = \frac{1}{5-1}\sum_{i=1}^{5}(\hat{B}_i - \bar{B})^2$$

The sampling error component of the error variance is computed using the estimates of the A and B parameters from the first plausible value with the total survey weight, and the set of estimates using the first plausible value with all of the replicate weights.[2] The sampling error components of the A and B parameters are computed as

$$S_A = \sum_{j=1}^{62}\left(\hat{A}_{1j} - \hat{A}_{1,j=total\ survey\ weight}\right)^2$$

$$S_B = \sum_{j=1}^{62}\left(\hat{B}_{1j} - \hat{B}_{1,j=total\ survey\ weight}\right)^2$$

The total error variances for the A and B parameters are obtained by combining the measurement and sampling error variances as

$$E_A = S_A + \left(1 + \frac{1}{5}\right) M_A$$

$$E_B = S_B + \left(1 + \frac{1}{5}\right) M_B$$

The square root of these error variances correspond to the jackknife standard errors for the intercept and slope parameter of the moderation linking.

[2] PIRLS has 75 replicates compared with 62 in NAEP. The results reported in this report use the first 62 of the 75 replicate weights from PIRLS.