2011 NAEP-TIMSS Linking Study:

Linking Methodologies and Their Evaluations

October 2013

U.S. Department of Education

NCES 2013-469

This NCES document was prepared to supplement the “Linking Methodologies” section presented in the Global Context report (NCES 2013-460).

2011 NAEP-TIMSS Linking Study: Linking Methodologies and Their Evaluations

TABLE OF CONTENTS

NAEP-TIMSS Linking Study 3
Linking Methodologies 3
LINKING METHODOLOGY: CALIBRATION 6
LINKING METHODOLOGY: PROJECTION 10
LINKING METHODOLOGY: STATISTICAL MODERATION 26
Evaluations of the Methodologies 53
EVALUATION DESIGN 53
PRIMARY FINDINGS AND CONCLUSIONS – STAGE 1 54
PRIMARY CONCLUSIONS – STAGE 2 78
Recommendations 96
References 98

This NCES document was prepared to supplement the "Linking Methodologies" section presented in the Global Context report (NCES 2013-460).

NAEP-TIMSS Linking Study

The 2011 NAEP-TIMSS linking study conducted by the National Center for Education Statistics (NCES) was designed to predict TIMSS scores for the U.S. states that participated in 2011 NAEP mathematics and science assessment of eight-grade students. The study design involved four samples of students:

1. Students assessed in NAEP mathematics or science during the winter (January-March) 2011 NAEP administration (**NAEP operational/national sample**);
2. Students in the United States assessed in TIMSS (mathematics and science) during the spring (April-June) 2011 TIMSS administration (**TIMSS U.S. operational/national sample**);
3. Students assessed during the 2011 NAEP testing window with booklets, referred to as braided booklets, containing one block of NAEP and one block of TIMSS items (which followed **NAEP administration procedures**); and
4. Students assessed during the spring 2011 TIMSS testing window with booklets, also referred to as braided booklets, containing one block of NAEP and three blocks of TIMSS items (which followed **TIMSS administration procedures**).

The braided-booklet sample under the NAEP administration window (i.e., sample 3) was given the NAEP-like booklets, which were designed to appear as similar as possible to a regular NAEP assessment booklet and were administered under the same conditions as NAEP. Similarly, the braided-booklet sample under the TIMSS administration window (i.e., sample 4) was given the TIMSS-like booklets. Those booklets were designed to appear as similar as possible to a regular TIMSS assessment booklet and were administered under nearly the same conditions as TIMSS. In addition, the braided booklets in the 2011 TIMSS window were administered in the same schools in which TIMSS was administered, with one intact classroom randomly assigned to the U.S. TIMSS national sample and another to the braided-booklet sample.

In addition to these linking study samples, nine states—Alabama, California, Colorado, Connecticut, Indiana, Florida, Massachusetts, Minnesota, and North Carolina—participated in 2011 TIMSS directly as separate jurisdictions and, therefore, received actual TIMSS scores. These nine states provided a “validation sample” upon which the NAEP-TIMSS link was evaluated. The validation states were selected based on their state enrollment and willingness to participate, and also on whether they as a whole represented a substantial range of performances relative to the national NAEP average, had previous experience as benchmarking participants in TIMSS, and were geographically diverse. See Figure 1 for details on sample sizes.

Linking Methodologies

The purpose of conducting the 2011 NAEP-TIMSS linking study was two-fold. The study was conducted to see whether it is possible to predict TIMSS scores for the states that did not

participate in the TIMSS assessment. Secondly, the study was conducted to identify a method among various methodologies suggested in the literature for linking two assessments that are somewhat different. Mislevy (1992) and Linn (1993) proposed a type of taxonomy in categorizing the linking methodologies into four forms—equating, calibration, projection, and moderation. Linking NAEP and TIMSS is an effort to link assessments based on different frameworks. It is clear that equating is not a feasible approach. (See Kolen & Brennan, 2004, for the requirements for equating.) The other three linking methods—moderation, projection, and calibration—were applied in linking NAEP and TIMSS assessments conducted in 2011. Among the three methods, calibration linking is appropriate when two assessments: (1) are based on the same frameworks but possess different test specifications and different statistical characteristics or (2) have frameworks that share common features and/or uses, but still are viewed as different and with different test specifications (Kolen & Brennan, 2004). On the other hand, the projection and moderation linking methods can be used without the expectation that "the same things" are being measured (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999). In addition, as will be discussed later in the paper, additional braided-booklet samples are required for the calibration and projection linking methods, but not the moderation method. The accuracy of the predicted TIMSS scores was evaluated by comparing the predicted and actual TIMSS scores for the nine validation states.

Based on the evaluation of the linking results, NCES has adopted the statistical moderation technique to report predicted TIMSS scores for the 43 U.S. states/jurisdictions that did not participate in the 2011 TIMSS grade 8 assessments at the state level. This decision was made because the evaluation of results showed that all three methods of linking yielded essentially the same predicted TIMSS results. In addition, among the three methods, the statistical moderation technique is the simplest method requiring the estimation of the fewest parameters (i.e., the means and standard deviations of the U.S. national public school samples for NAEP and TIMSS). The method also could be applied to the extant national samples of NAEP and TIMSS and did not require the use of separate braided-booklet samples that were required for the calibration and projection methods of linking. This implies that NCES has the option of conducting future NAEP-TIMSS linking studies using statistical moderation without the additional resources needed for the braided-booklet samples. Selecting a relatively simple and efficient methodology allows NCES to conduct additional linking studies in the future.

Multiple NCES contractors were involved in carrying out the linking study. One NCES contractor, Educational Testing Service (ETS), applied the calibration and the statistical projection methods, while another, American Institutes for Research (AIR), applied the statistical moderation method. In the next section of this paper, descriptions of the methods applied in the 2011 linking study are presented. A third contractor, the Human Resources Research Organization (HumRRO), evaluated the results obtained by the three linking methods and made a set of recommendations based on their evaluation. The linking results and the

recommendations were discussed with various expert panels, namely, the NAEP Design and Analysis Committee and the National Assessment Governing Board. HumRRO's evaluation of the linking results and their recommendations are presented in the final section of this paper.

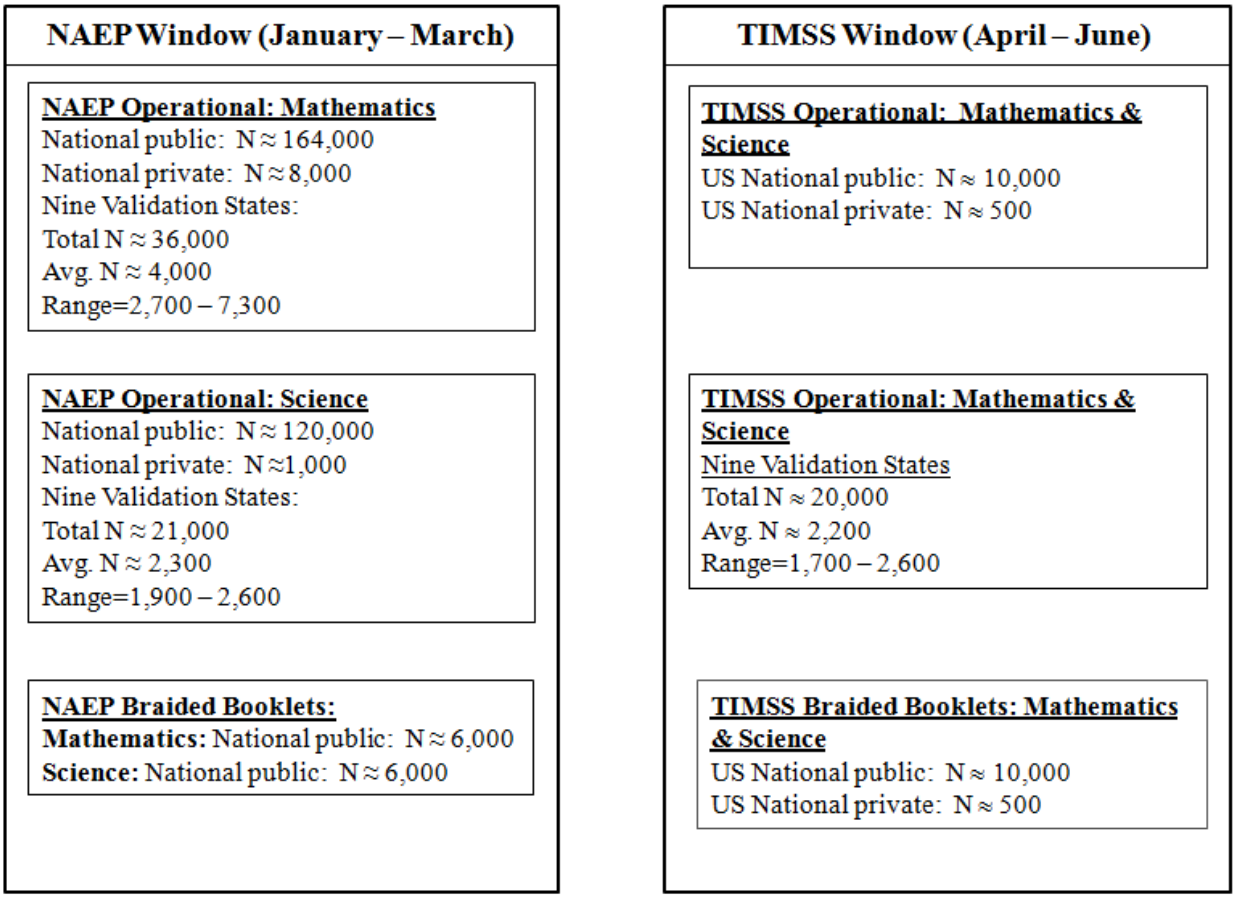


Figure 1. Sample sizes for the linking study

Linking Methodology: Calibration

In the literature, the term *calibration* has several different meanings and connotations. We use it here to refer to a procedure of putting all the NAEP and TIMSS items in a given domain (mathematics or science) on a common item response theory (IRT) scale. As discussed in Kolen and Brennan (2004, page 430), calibration linking is a type of linking used when the two assessments are based on

(1) the same framework but different test specifications and different statistical characteristics, or

(2) different frameworks and different test specifications, but the frameworks are viewed as sharing common features and/or uses.

Calibration linking is typically used in a nonequivalent groups anchor test (NEAT) design in which a set of "common items" or common test questions is administered to all groups. For instance, student sample 1 is administered item sets A and B, while student sample 2 is administered item sets B and C. Items in set B are the common items. Although NAEP and TIMSS are based on different frameworks and have different test specifications, the two assessments do share a number of common features (Neidorf, Binkley, Gattis, & Nohara, 2006, Nohara, 2001, Provasnik et al., 2012). Therefore, calibration linking is used based on the second type of linking condition listed above.

As shown in Figure 1, the 2011 NAEP-TIMSS linking study design included braided-booklet samples that took items from both NAEP and TIMSS at the same time and under the same testing conditions. Consequently, NAEP items were common among the 2011 operational NAEP sample and the two braided-booklet samples (one in the NAEP administration window, and the other in the TIMSS administration window), and TIMSS items were common among the 2011 operational TIMSS U.S. sample and the two braided-booklet samples. Figure 2 illustrates how the study design provided common items in linking NAEP and TIMSS. The study thus supports the use of calibration linking, the goal of which was to express the IRT item parameters for the 2011 NAEP items on the TIMSS scale.

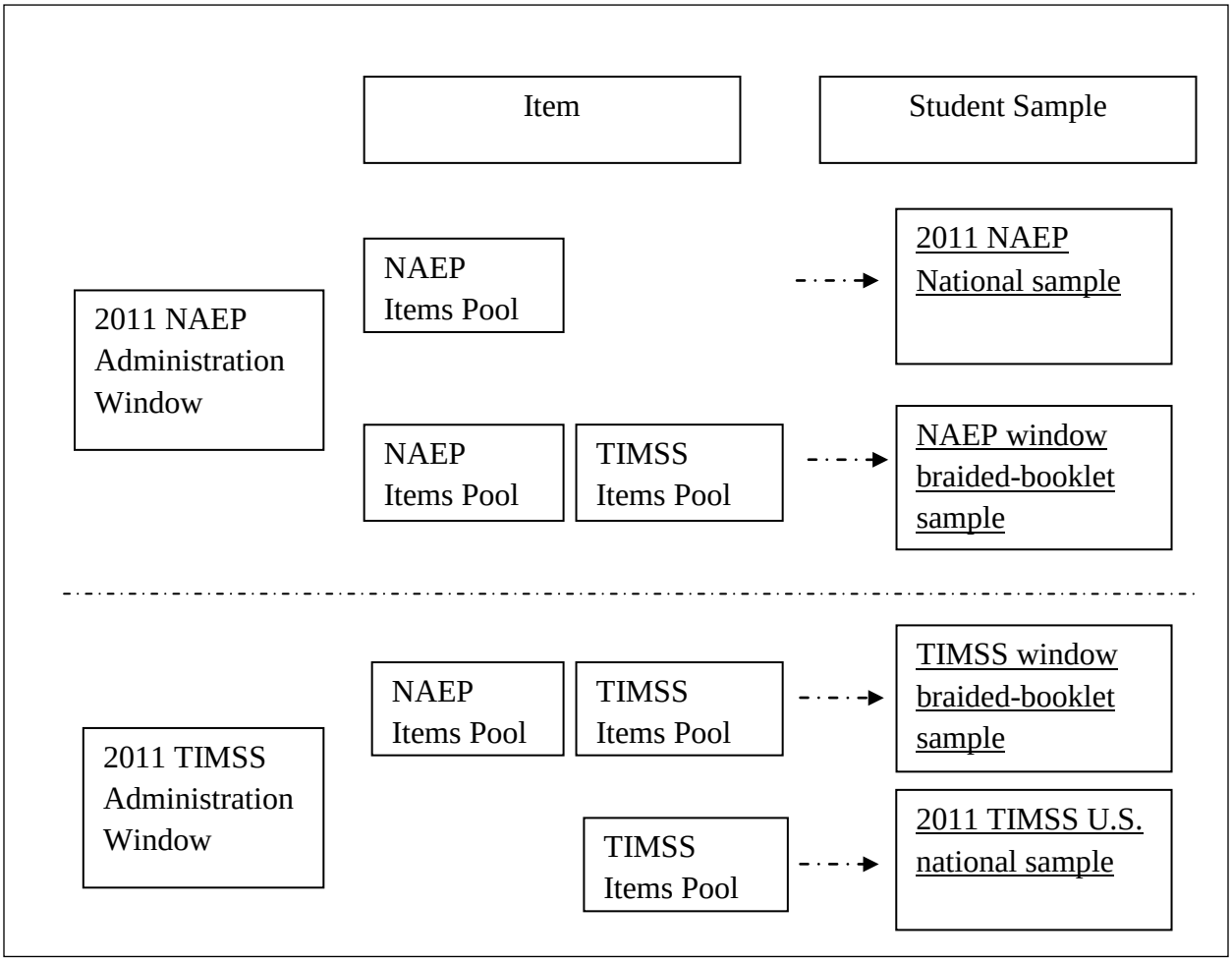


Figure 2: *Study Design of the 2011 NAEP-TIMSS Linking Study*

The objective of the NAEP-TIMSS linking study was to use states' 2011 NAEP scores to predict their mean TIMSS scores and percentages of students reaching each of the TIMSS international benchmark levels. Therefore, we wanted the predicted TIMSS scores to be placed on the existing TIMSS scale, which was established based on countries that participated in TIMSS (Foy, Brossman, & Galia, 2012). Consequently, for the calibration linking analysis, we employed the fixed parameter calibration method. That is, we first fixed the IRT item parameters for the TIMSS items at their values from the TIMSS 2011 operational analysis. Next, the items from the NAEP assessment were placed onto the established TIMSS scale by calibrating the items from the NAEP and TIMSS assessments together but keeping TIMSS item parameters fixed.

Three major steps were involved in the fixed parameter calibration linking: (1) calibrating the NAEP items onto the TIMSS scale; (2) estimating population proficiency scores in TIMSS for the 2011 NAEP samples in mathematics and science; and (3) placing the predicted proficiency scores on the metrics used to report TIMSS results. In the following sections, we describe each step of the calibration linking analysis.

Step 1: **Calibrating the NAEP items onto the TIMSS scale**

For this first step, we used the item parameters for the eighth-grade TIMSS mathematics and science items from the TIMSS 2011 operational analysis. In the TIMSS operational analysis, the two IRT scales, one for mathematics and the other for science, were constructed separately. In linking assessments between administrations, TIMSS uses concurrent calibration, which calibrates item parameters for the items in the current assessment through a concurrent calibration of the data from the current assessment and from the previous assessment (See, for example, Foy, Brossman, & Galia, 2012, for details).

In line with TIMSS operational practice, we conducted two separate fixed parameter calibrations, one for mathematics and the other for science. The item parameters of the TIMSS items were fixed at the values obtained from the TIMSS 2011 operational analysis, and the NAEP item parameters were calibrated) onto the TIMSS IRT scale. The item responses from three groups of students—the 2011 NAEP national sample[1], the NAEP window braided-booklet sample, and the TIMSS window braided-booklet sample—were used in the calibration, and the proficiency distributions for the three groups were not constrained to be equal. Note that the 2011 TIMSS sample was not included. This is because only the NAEP item parameters need to be estimated in the fixed parameter calibration; no NAEP items were administered to the 2011 TIMSS sample.

For dichotomously scored items, two- and three-parameter logistic models (Lord & Novick, 1968) were used, while for polytomously scored items the generalized partial-credit model (Muraki, 1992) was used. Details about the IRT model fit evaluation and the estimated item parameters for all 2011 NAEP mathematics and science items from fixed parameter calibration will be provided in the forthcoming NAEP-TIMSS linking study technical report.

Step 2: **Estimating population proficiencies in TIMSS for the 2011 NAEP national sample**

In the second step, we took the IRT item parameters for the NAEP items estimated in the first step and employed a procedure called “conditioning” to estimate mathematics and science proficiency distributions for the 2011 NAEP national sample[2]. The item parameters estimated in step 1 served the purpose of setting the TIMSS IRT scale on which these proficiencies were estimated. Plausible values—random draws from the predictive scale score distribution for each respondent on the TIMSS IRT scale (see von Davier, Gonzalez, & Mislevy, 2009) were generated for all students in the 2011 NAEP national sample. The plausible values were used to estimate student subgroup proficiencies and associated variances. We drew 20 plausible values per respondent in the 2011 NAEP national sample.

[1] The 2011 NAEP national sample included students from both public and private schools.
[2] Full descriptions of the conditioning procedure can be found in Beaton, 1987; Mislevy, Beaton, Kaplan, & Sheehan, 1992; and Mislevy, Johnson, & Muraki, 1992.

Step 3: **Transform the predicted proficiency distributions for the 2011 NAEP national samples to the TIMSS reporting metrics**

The third step was to transform the proficiency distributions obtained in step 2 from the TIMSS IRT scales to the TIMSS scale score reporting metrics. A mean-sigma linear transformation procedure was applied that transformed the distribution of the 2011 NAEP national sample from the TIMSS IRT scale to match the mean and standard deviation of the proficiency distribution of the 2011 TIMSS U.S. national sample that was available on the TIMSS reporting metric. The transformation was carried out separately for mathematics and science. Student plausible values were used in computing the means and standard deviations of the score distribution. The transformation equation was as follows:

$$PV_{\text{Target}} = \hat{A} \bullet PV_{\text{Calibrated}} + \hat{B} \qquad \text{(E1)}$$

where

- $PV_{\text{Calibrated}}$ was the plausible value on TIMSS IRT scale from fixed parameter calibration;
- PV_{Target} was the plausible value on the TIMSS reporting metric, obtained using linear transformation parameter estimates $\hat{A}$ and $\hat{B}$

  $\hat{A} = SD_{\text{target}} / SD_{\text{calibrated}}$,

  $\hat{B} = M_{\text{target}} - \hat{A} \bullet M_{\text{calibrated}}$;

- SD_{target} = the estimated standard deviation of the proficiency distribution for the 2011 TIMSS U.S. national sample on the TIMSS reporting metric;
- $SD_{\text{calibrated}}$ = the estimated standard deviation of the proficiency distribution for the 2011 NAEP national sample on the TIMSS IRT metric;
- M_{target} = the estimated mean of the proficiency distribution for the 2011 TIMSS U.S. national sample on the TIMSS reporting metric; and
- $M_{\text{calibrated}}$ = the estimated mean of the proficiency distribution for the 2011 NAEP national sample on the TIMSS IRT metric.

The estimated transformation parameters are listed in Table 1.

Table 1: Estimated Transformation Parameters for Achievement Scores for Calibration Linking: 2011

	Parameter Estimates	
Mathematics	$\hat{A}$	$\hat{B}$
	106.999	484.485
	Parameter Estimates	
Science	$\hat{A}$	$\hat{B}$
	106.666	495.330

Linking Methodology: Projection

Conceptually, projection is a type of statistical machinery that estimates a relationship between scores on two tests, and then derives predictions ("projections") of scores on one test from scores on the other test (Mislevy, 1992). Projection linking can be applied without the assumption or expectation that the same constructs are being measured by the two tests (Feuer et al., 1999). Projection linking is directional. That is, projecting NAEP scores onto the TIMSS scale is different from projecting TIMSS scores onto the NAEP scale. In addition, this approach requires a linking sample where (groups of) students take items from both tests. The projection linking analysis uses the linking sample to model the relationships between scores on the two assessments.

In this linking study, the students in the braided-booklet samples provided answers to test questions or items from both NAEP and TIMSS at the same time and under the same conditions, without knowing whether they were given an operational test booklet, or a braided booklet with items from two different assessments.

In addition to responding to cognitive test items, the braided-booklet samples assessed during the NAEP administration window were given the NAEP survey questionnaires. Likewise, the braided-booklet sample under the TIMSS administration window took the TIMSS survey questionnaires. Therefore, the current design allowed us to directly estimate the joint NAEP-TIMSS population-structure model by using survey questionnaires and students' responses to the cognitive test questions and taking into account the relationship between the two assessments. The conditional proficiency distribution of TIMSS given the NAEP proficiency distribution can subsequently be derived from the braided-booklet sample and serve as the projection linking function.

Given the availability of the braided-booklet samples under both NAEP and TIMSS administration windows as shown in Figure 1, we were able to derive two projection functions for each subject domain and compare them for consistency. Note that in theory, the braided-booklet samples from both administration windows can be combined to estimate a single projection function for each subject. However, as will be more evident from the description of the projection linking procedure that follows, forming a single projection function would not have been a straightforward replication of deriving a projection function for an individual braided-booklet sample, as the students in the NAEP window took only either mathematics or science, while those in the TIMSS window took test items from both subjects. For this study, the braided-booklet samples across assessment windows were not combined in deriving projection functions. Next, a six-step procedure, which was applied to carry out the projection linking, is described.

Step 1: **Apply the NAEP and TIMSS latent proficiency scale item IRT parameters to the linking sample item responses**

The NAEP and TIMSS latent proficiency scales are both estimated based on a combination of IRT models (see, for example, Allen, Donoghue, & Schoeps, 2001, Foy, Galia & Li, 2008). For dichotomously scored items two- and three-parameter logistic models (Lord & Novick, 1968) were used while for polytomously scored items the generalized partial-credit model (Muraki, 1992) was used.

The braided instrument that was administered to the braided-booklet samples included the complete pool of items administered in the 2011 NAEP and TIMSS mathematics and science assessments. We used the operational 2011 NAEP item parameter estimates[3] to calculate NAEP proficiency estimates for the braided-booklet samples. Likewise, we applied the operational 2011 TIMSS item parameter estimates from the overall mathematics and science scales in the calculation of TIMSS proficiency estimates. Details about the IRT model fit evaluation will be provided in the forthcoming NAEP-TIMSS linking study technical report.

Step 2: **Estimate the projection function for the braided-booklet samples**

In the second step, the “conditioning” procedure was employed to estimate the joint NAEP and TIMSS proficiency distribution through a latent regression model, based on the IRT parameters from step 1, student responses to the subset of items they received, as well as other relevant and available background information[4]. For the mathematics linking sample in the NAEP

[3] For 2011 NAEP science, an overall univariate IRT scale was established in the operational analysis with the IRT model item parameters estimated for each item on that scale. Those item parameter estimates were applied directly to the braided-booklet samples. For 2011 NAEP mathematics, five separate IRT latent scales were constructed in the operational analysis, one for each content domain. For the purpose of this linking study, an overall univariate scale was first established for 2011 NAEP mathematics and linked to the NAEP mathematics reporting scale. The IRT model item parameters were estimated for each item on that overall scale, which were then applied to the braided-booklet samples.

[4] Full descriptions of the conditioning procedure can be found in Beaton, 1987; Mislevy, Beaton, Kaplan, & Sheehan, 1992; and Mislevy, Johnson, & Muraki, 1992.

administration window, a bivariate latent regression population-structure model was used to estimate this joint distribution of NAEP and TIMSS mathematics scores. Plausible values were generated for all students in the braided-booklet sample. These plausible values can subsequently be used to represent probabilities in joint and conditional proficiency distributions, and allow unbiased group-level estimates. Similar to the calibration method, 20 plausible values were drawn for individual students in the braided-booklet sample.

The same conditioning procedures were used to estimate the joint distribution of NAEP and TIMSS science proficiencies from the science linking sample in the NAEP administration window. Students in the TIMSS window linking sample were administered items from both subjects (mathematics and science) and assessments (NAEP and TIMSS), and so a four-variate latent regression was conducted where each combination of subject and assessment comprised a dimension—NAEP mathematics, NAEP science, TIMSS mathematics, and TIMSS science.

Step 3: **Transform the proficiency distributions for the braided-booklet samples from the IRT metrics to the reporting metrics**

The NAEP and TIMSS proficiency distributions for the braided-booklet samples obtained from step 2 were estimated on the NAEP and TIMSS IRT scales, respectively. The third step is to place the proficiency distributions on the NAEP and TIMSS reporting metrics.

Both NAEP and TIMSS apply linear transformation to transform results from IRT metrics to the appropriate reporting metrics. In operational TIMSS analysis, based on concurrent IRT calibration approaches, linear transformation parameters are estimated that transform the distribution of the previous assessment data under the concurrent calibration to match means and standard deviations of the distribution of these data that are available on the reporting metric. Those transformation parameter estimates are then used to place the current assessment data on the TIMSS reporting scale. Student plausible values are used in computing the means and standard deviations of the score distribution. There exist five plausible values for individual students. A total of five sets of transformation parameter estimates ($\hat{A}_i$'s and $\hat{B}_i$'s) are available, one for each plausible value. The transformation equation is as follows:

$$PV_{\text{i,Target}} = \hat{A}_i \bullet PV_{\text{i,Calibrated}} + \hat{B}_i \qquad \text{(E2)}$$

where

- $PV_{\text{i,Target}}$ was the plausible value *i* on the transformed TIMSS reporting scale;
- $PV_{\text{i,Calibrated}}$ was the plausible value i on the original IRT scale on the TIMSS IRT scale; and
- $\hat{A}_i$ and $\hat{B}_i$ were the estimates of the linear transformation parameters.

Instead of obtaining and applying five sets of transformation parameter estimates, NAEP estimates only one set of transformation parameters $\hat{A}$ and $\hat{B}$, which is computed by first averaging the means and standard deviations of the score distribution obtained from both the IRT and NAEP reporting metrics.

For the braided-booklet samples in the NAEP-TIMSS linking study, given that the original 2011 NAEP item parameter estimates were used in estimating the plausible values on the calibration scale, we applied the transformation parameter estimates $\hat{A}$ and $\hat{B}$ from the operational 2011 NAEP analysis[5] to place the NAEP plausible values on the NAEP reporting metric. Likewise, the transformation parameter estimates from the operational 2011 TIMSS analysis were used to place the TIMSS plausible values from the IRT scale on the TIMSS reporting metric. To transform 20 plausible values drawn in step 2 to the TIMSS reporting metrics, each of the five sets of transformation parameter estimates from the operational 2011 TIMSS analysis was applied to four different plausible values.

Step 4: **Smooth the projection functions from the braided-booklet samples**

Taking the NAEP and TIMSS plausible values obtained in step 3, the joint NAEP-TIMSS proficiency distribution for each subject estimated from the plausible values was smoothed using a continuous bivariate exponential family distribution (Haberman, 2011). With the NAEP and TIMSS latent proficiencies presented as a joint continuous distribution, the projection function was smoothed by deriving the conditional distribution of TIMSS proficiency given NAEP proficiency.

Step 5: **Predict TIMSS scores for all the states**

The prediction functions derived in step 4 were used to predict TIMSS plausible scores for students in the 2011 NAEP national sample. For each subject, mathematics and science, there were five NAEP plausible values available for each student in the 2011 NAEP national sample. Four plausible values were drawn from the conditional TIMSS proficiency distribution for each given NAEP plausible value. Then, for each student, a total of 20 new sets of predicted TIMSS plausible values were drawn. The predicted TIMSS plausible values were used to estimate individual state average TIMSS scores and the percentage of students reaching each of the TIMSS international benchmarks.

[5] For 2011 NAEP science, an overall univariate scale was established in the operational analysis. Therefore the transformation constants A and B from the operational 2011 NAEP science analysis were directly applied. For 2011 NAEP mathematics, five separate scales were constructed in the operational analysis, one for each content domain. For the purpose of this linking study, an overall univariate scale was first established for 2011 NAEP mathematics and linked to the NAEP mathematics reporting scale. The transformation constants A and B obtained from the overall NAEP mathematics scale were applied to the braided-booklet samples.

Step 6: **Additional linear adjustment to the predicted overall TIMSS mathematics and science distributions**

The predicted TIMSS plausible values obtained from step 5 of the projection linking procedure are estimates of how students in the 2011 NAEP sample would have performed if they had taken TIMSS, to the extent that differences between NAEP and TIMSS are accounted for in the projection functions. To better facilitate comparisons to other countries and subnational education systems that participated in TIMSS2011 during the TIMSS window and under TIMSS administration conditions, the distributions of predicted TIMSS plausible values from the 2011 NAEP national sample were then aligned (through a linear transformation adjustment) to the distribution of TIMSS plausible values from the 2011 TIMSS U.S. national sample, separately for mathematics and science.

$$PV_{\text{Target_adjusted}} = \hat{A} \bullet PV_{\text{Target}} + \hat{B} \qquad \text{(E3)}$$

Where

- PV_{Target} was the plausible value on the TIMSS reporting scale from step 5 of projection linking;
- $PV_{\text{Target_adjusted}}$ was the plausible value on the TIMSS reporting scale after the linear adjustment, both for the 2011 NAEP assessment; and
- $\hat{A}$ and $\hat{B}$ were the estimates of the adjustment function parameters

  $\hat{A} = SD_{\text{target_adjusted}} / SD_{\text{target}}$,

  $\hat{B} = M_{\text{target_adjusted}} - \hat{A} \bullet M_{\text{calibrated}}$.

Table 2 contains the estimates of the adjustment function parameters, separately for mathematics and science, and for the different projection functions obtained from the NAEP and TIMSS window braided-booklet samples.

Table 2: Estimated Linear Adjustment Function Parameters for Achievement Scores for Projection Linking: 2011

Projection with NAEP Window Braided-booklet Sample		
Mathematics	Parameter Estimates	
	$\hat{A}$	$\hat{B}$
	0.937	34.336
Science	Parameter Estimates	
	$\hat{A}$	$\hat{B}$
	0.984	9.298
Projection with TIMSS Window Braided-booklet Sample		
	Parameter Estimates	
Mathematics	$\hat{A}$	$\hat{B}$
	0.906	51.929
	Parameter Estimates	
Science	$\hat{A}$	$\hat{B}$
	0.917	62.789

Findings from calibration and statistical projection

The key findings from the calibration and projection linking methods are presented next. For projection linking, as discussed above, two separate projection functions were developed for each subject—one using the braided-booklet sample data from the NAEP testing window and one using that from the TIMSS testing window. Besides the main goal of providing predicted TIMSS results for the states that took NAEP, another question of interest in the study is whether the braided-booklet samples and instruments that were developed for the two assessment windows were necessary for carrying out a projection type of linkage. Differences were found between the projection functions obtained from the two linking samples. For the nine validation states that had their actual TIMSS scores, before applying the linear adjustment as described in step 6 of the projection linking, the projected state TIMSS means were closer to their actual results when using the projection function derived from the braided-booklet sample from the NAEP testing window. The linear adjustment applied to the projection-based TIMSS proficiency distribution generally reduced differences between the predicted and actual state TIMSS results. In addition, after incorporating the linear adjustment, the projection-based results with projection functions derived from the two testing windows were comparable. Details on the projection-

based results with and without the linear adjustments will be provided in the forthcoming NAEP-TIMSS linking study technical report. For the purpose of comparing the predicted TIMSS results obtained from different linking approaches, we use the projection-based results, incorporating the linear adjustment, derived from the NAEP window braided-booklet sample.

Tables 3a and 3b contain the state-level predicted TIMSS mean scores from both calibration and projection linking approaches and differences thereof for Mathematics (Table 3a) and Science (Table 3b). In addition, the actual TIMSS mean scores for the validation states obtained by directly participating in TIMSS are presented along with differences between those and the predicted scores. The last column shows that the two linking approaches result in largely comparable predicted results. Based on the nine validation states, it shows that the calibration has a very slight edge over projection when compared to the actual TIMSS results. That being said, it is also observed from columns 4 and 6 that there were sizeable discrepancies between predicted and actual state results for more than half of the validation states.

The complete set of predicted TIMSS results, including predicted state-level means and percentages of students at or above the four TIMSS international benchmarks are listed in Tables 4a and 4b. These four benchmarks are: 625 (Advanced), 550 (High), 475 (Intermediate), and 400 (Low). These benchmarks provide a way to interpret the average scores and understand how students' proficiency in mathematics and science varies along the TIMSS scale.

Table 3a: Predicted and (for validation states) actual TIMSS state mean Mathematics scores

State	Actual TIMSS State Mean Math Score	Predicted TIMSS State Mean Math Score				Difference Between Predictions (Calibration - Projection)
		Calibration Linking		Projection Linking		
		Estimate	Residual	Estimate	Residual	
Alabama	466	478	12	480	14	-2
California	493	486	-7	487	-5	-1
Colorado	518	526	8	525	8	1
Connecticut	518	516	-1	516	-2	1
Florida	513	496	-17	497	-17	0
Indiana	522	513	-9	512	-9	0
Massachusetts	561	540	-20	538	-22	2
Minnesota	545	533	-12	532	-13	2
North Carolina	537	515	-22	514	-23	1

NOTE: the numbers in the last column “Difference Between Predictions” may differ from the calibration linking estimate minus the projection linking estimate due to rounding.

Table 3b: Predicted and (for validation states) actual TIMSS state mean Science scores

State	Actual TIMSS State Mean Science Score	Predicted TIMSS State Mean Science Score				Difference Between Predictions (Calibration - Projection)
		Calibration Linking		Projection Linking		
		Estimate	Residual	Estimate	Residual	
Alabama	485	497	11	500	15	-4
California	499	498	0	500	1	-2
Colorado	542	546	4	544	2	2
Connecticut	532	532	0	531	0	0
Florida	530	517	-13	518	-12	-1
Indiana	533	527	-6	527	-6	0
Massachusetts	567	547	-19	545	-22	2
Minnesota	553	546	-7	544	-9	2
North Carolina	532	515	-17	516	-15	-2

NOTE: the numbers in the last column “Difference Between Predictions” may differ from the calibration linking estimate minus the projection linking estimate due to rounding.

Table 4a: Predicted TIMSS state means and benchmark percentages from calibration and projection linking, Mathematics

State	Calibration Linking										Projection Linking									
	Mean	SE	>=400		>=475		>=550		>=625		Mean	SE	>=400		>=475		>=550		>=625	
			Pct	SE	Pct	SE	Pct	SE	Pct	SE			Pct	SE	Pct	SE	Pct	SE	Pct	SE
Alabama	478	4.0	84	1.6	54	2.2	17	1.7	2	0.9	480	3.7	85	1.4	54	2.1	18	1.5	2	0.5
California	486	3.5	85	1.1	56	1.6	22	1.3	5	0.7	487	3.4	85	1.2	57	1.9	23	2.0	4	0.7
Colorado	526	3.5	95	1.1	76	1.5	39	1.9	9	1.2	525	3.5	95	0.7	76	1.5	39	2.4	8	1.5
Connecticut	516	3.6	94	1.0	71	2.2	34	1.9	7	1.2	516	3.7	93	0.8	71	1.8	34	1.7	7	1.1
Florida	496	3.2	90	1.2	62	1.8	24	1.7	4	0.6	497	3.2	89	1.1	62	1.7	25	1.4	4	0.6
Indiana	513	3.4	94	0.8	71	1.7	31	2.0	5	0.8	512	3.2	94	0.9	71	1.7	31	1.6	5	1.0
Massachusetts	540	3.3	96	0.6	82	1.5	46	2.2	11	1.2	538	3.3	96	0.7	81	1.7	46	2.0	11	1.4
Minnesota	533	3.3	95	0.6	80	1.4	43	2.1	10	1.5	532	3.4	95	0.7	79	1.6	42	1.7	9	1.3
North Carolina	515	3.5	93	1.5	70	1.9	33	1.9	7	1.3	514	3.4	93	1.1	70	2.0	33	1.8	7	1.0

Table 4b: Predicted TIMSS state means and benchmark percentages from calibration and projection linking, Science

State	Calibration Linking										Projection Linking									
	Mean	SE	>=400		>=475		>=550		>=625		Mean	SE	>=400		>=475		>=550		>=625	
			Pct	SE	Pct	SE	Pct	SE	Pct	SE			Pct	SE	Pct	SE	Pct	SE	Pct	SE
Alabama	497	3.9	87	1.4	64	2.0	27	2.0	4	1.0	500	3.8	88	1.3	65	2.0	29	1.9	5	0.9
California	498	3.7	86	1.2	63	2.0	29	1.8	6	0.8	500	3.7	87	1.5	64	2.0	30	1.7	7	1.0
Colorado	546	3.9	96	1.1	82	1.8	51	2.5	15	1.8	544	3.7	96	0.7	82	1.4	49	1.8	14	1.8
Connecticut	532	3.5	94	0.9	77	1.7	44	2.1	11	1.4	531	3.5	94	0.8	77	1.7	43	2.1	11	1.2
Florida	517	3.5	91	1.2	71	1.8	37	2.6	8	0.9	518	3.5	92	1.2	72	1.8	37	2.2	8	0.8
Indiana	527	3.1	94	1.0	77	1.5	42	2.0	8	1.0	527	3.2	94	1.1	76	1.8	41	1.7	9	0.9
Massachusetts	547	3.3	95	0.7	83	1.4	53	1.7	16	1.2	545	3.4	95	0.8	82	1.3	51	1.9	15	1.6
Minnesota	546	3.3	96	0.9	84	1.3	52	1.8	13	1.4	544	3.4	96	0.7	83	1.2	50	2.0	13	1.3
North Carolina	515	3.4	92	1.8	71	1.6	35	1.8	7	0.9	516	3.4	92	1.1	72	2.2	35	1.8	8	1.0

Standard Error Estimation For Calibration and Projection

The 2011 TIMSS eighth-grade achievement results for the participating countries, subnational education systems, and the nine states in the United States, were released in December 2012. The standard errors of the actual TIMSS mean scores and benchmark percentages include sampling and measurement components

$$Var = Var_{sampling} + Var_{measurement} \quad \text{(E4)}$$

As a result of the linking study, we predicted TIMSS state results for the states that participated in NAEP. For all the states (validation plus non validation states), the error variance associated with predicted TIMSS results can be expressed as

$$Var = Var_{sampling} + Var_{linking} + Var_{measurement} \quad \text{(E5)}$$

In both the NAEP and TIMSS assessments, a jackknife procedure is used to calculate sampling error for any reporting statistic directly. As discussed before, for calibration linking, the predicted TIMSS proficiency distribution for the NAEP national sample was transformed from IRT scale to the TIMSS reporting metric. For projection linking, the predicted TIMSS proficiency distribution for the NAEP national sample obtained was adjusted to have the same mean and standard deviation as the reported TIMSS U.S. national sample. It can be conjectured that the transformation/adjustment function parameter estimates from calibration and projection linking are subject to non-negligible error. Therefore, a jackknife procedure was employed at the transformation/adjustment stage as well as the summary statistics estimation stage to estimate both sampling and linking errors. The standard errors for the predicted state-level TIMSS results are provided in Tables 4a and 4b.

The linking study can be thought of as a linking and prediction question where state-level TIMSS results are to be predicted. The variance estimated in equation (E5) captures the uncertainty of the linking function. However, there is also uncertainty associated with predicting a new point based on the linking function, which is referred to as prediction residual error variance. How to estimate prediction residual error variance could be challenging, given that a number of factors are involved that (a) may not be separable and (b) may represent not only random variance, but also bias. The mean squared error (MSE) can be used to quantify the discrepancies between actual and predicted values (see equations (E6) and (E7) for the formula of MSE). The MSE of prediction includes both variance and bias squared. For example, in sampling, the variance would be random error due to drawing different samples. Bias would be the result of using different sampling rules (such as eligibility requirements) where the

populations used to draw the samples are no longer the same. The square of this (systematic) bias plus the (random) variance is the MSE.

In comparing the predicted TIMSS scores with the actual scores, the bias portion can be expected to be considerable due to the many differences in administration policies and procedures. Subsequently, treating MSE as the prediction residual error variance in standard error calculation and hypothesis testing might result in misleading statements, indicating no significant differences when there are real differences if results from equivalent samples and under equivalent conditions would have been compared. Yet, the MSE may give an indication of how large this combined error is relative to the three random error components discussed above.

Effort has been made to adjust the predicted state results with the intention to (partially) remove bias and to review the impact of factors related to differential accommodations and exclusions. The idea is that if all or most of the bias can be accounted for, the remaining term is a prediction random error term that can be used in hypothesis testing. The section below on Selection Bias provides an account of this effort so far and the following conclusion is drawn from this (preliminary) work. While some impact was detected, these corrections are ad-hoc and experimental in nature, do not fully account for many other sources of bias, and still need to be further studied in terms of removing bias components appropriately. But such analyses are insightful to assess what level of bias reduction could be obtained by applying some initial approaches.

Selection Bias and Predicted TIMSS Score Adjustments

To further evaluate the predicted state results, we define *prediction residual error* as the difference between predicted and actual state results on TIMSS, then *predicted residual sum of squares*, or PRESS, across the nine validation states can be used as a summary measure of the prediction model

$$PRESS = \sum_{i=1}^{9} \left(\hat{t}_i - t_i \right)^2 \qquad \text{(E6)}$$

where t_i is the actual observed state result for the ith validation state, and $\hat{t}_i$ is the predicted value. We further define MSE as

$$MSE_prediction = \frac{PRESS}{9} - \frac{\sum_{i=1}^{9} Var(\hat{t}_i)}{9} - \frac{\sum_{i=1}^{9} Var(t_i)}{9} \qquad \text{(E7)}$$

where $Var(\hat{t}_i)$ is the variance of the predicted result for the *ith* validation state, and $Var(t_i)$ is the variance of the actual result for the *ith* validation state. Taking calibration linking results as an example, the PRESS and MSE for the predicted mean scores from calibration linking were

computed across the nine validation states. The results are listed in the first row of Tables 6a and 6b.

To the extent that these discrepancies showed a consistent pattern, several possible factors were considered, including construct differences, administration differences, and sample/target population differences. Among those, a significant factor is the difference in exclusion rate/accommodation policy. As shown in Figure 3, TIMSS exclusion rates are in general higher than in NAEP, at the national level, and for individual validation states, largely because accommodations are not offered in TIMSS. Such difference in the selection of assessment samples is referred to as sample selection bias.

Two types of ad hoc adjustments were considered to assess and quantify the impact of selection bias due to differences in exclusion rates and accommodation policies. The first is to adjust the state exclusion rates in NAEP to be the same as in TIMSS. Note that we only know the exclusion rate for TIMSS at the state level for the validation states and, therefore, the following analyses are based on that subset only. With no information on which and how student groups are excluded in TIMSS but included in NAEP, this procedure presumed that those students that would most likely be excluded from TIMSS are the lowest performing accommodated (i.e., Students with Disabilities (SD) and/or English Language Learners (ELL)) students in NAEP. From each state sample, the exact number of accommodated students were identified and excluded such that NAEP state-specific "inclusion" rates matched TIMSS state-specific inclusion rates. The predicted state results were then computed based on the reduced NAEP state samples.

A second possible ad hoc adjustment would be to account for as many as possible bias factors and using a "residual" MSE as a fourth variance component (in addition to measurement, sampling, and linking variances) in standard error estimation. As described in the section, Evaluation of Methodologies, for the nine validation states the *prediction residual error* is negatively correlated with the state percentage of accommodation rates in NAEP. Subsequently, a simple linear regression was built and estimated to minimize the variance of *prediction residual error* for the nine states. The scores were adjusted before calculating the MSE. This approach is reasonable in principle. However, MSE contains estimation bias as well as variability. Given the limited state-level data about TIMSS' exclusion rates of SD and ELL students, it cannot be tested whether a sufficient amount of bias has been accounted for. In other words, it is not determinable whether a mostly random variance component is obtained or major sources of biases still are left unaccounted for MSE.

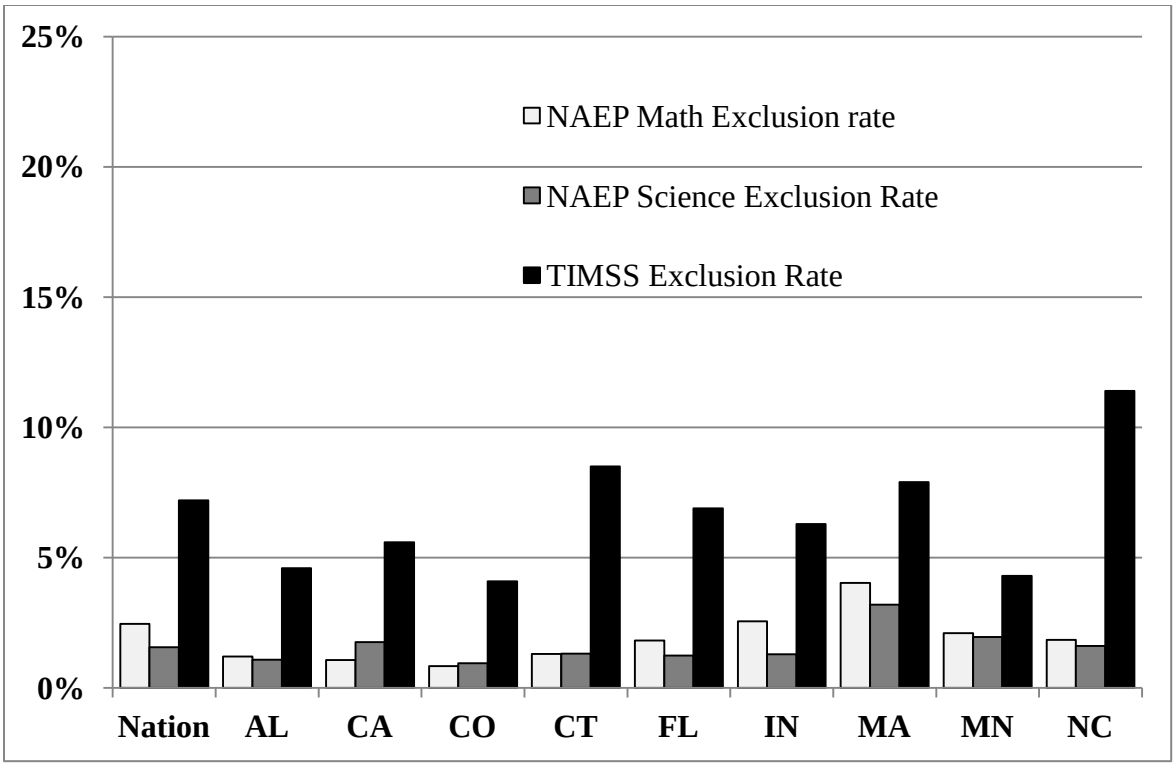


Figure 3: Percentage of students excluded from NAEP and TIMSS assessments at grade 8: 2011

Table 5a provides the actual TIMSS means and rankings of the nine validation states in mathematics. Also provided are rankings of the validation states and the *prediction residual errors* based on

a. Predicted TIMSS from the calibration linking (i.e., baseline); and

b. Predicted TIMSS from the calibration linking adjusted for exclusion rate differences between NAEP and TIMSS (i.e., reduced NAEP samples).

For reference, the ranking of the nine states based on the reported NAEP mathematics scores are listed as well. The PRESS and MSE computed from equations (E6) and (E7) are presented in Table 6a. Comparing to the predicted state means from calibration linking in the top row, the adjustment yields smaller *prediction residual errors* for most of the validation states and commensurate reduced PRESS and MSE values. Science results show similar patterns and are presented in Tables 5b and 6b.

Table 5a: Actual TIMSS means and predicted means and prediction residual errors for the nine validation states from calibration linking: Mathematics

Jurisdiction	Actual TIMSS Math			(a) TIMSS Math Predicted			(b) Predicted w/ Exclusion Rate Matching		Rank in 2011 NAEP Math
	Rank	Mean	SE	Rank	Residual	SE	Rank	Residual	
U.S. National		509	2.6						
Massachusetts	1	561	5.3	1	-20	3.3	1	-19	1
Minnesota	2	545	4.6	2	-12	3.3	2	-13	2
North Carolina	3	537	6.8	5	-22	3.5	4	-17	5
Indiana	4	522	5.1	6	-9	3.4	6	-10	6
Colorado	5	518	4.9	3	8	3.5	3	8	3
Connecticut	6	518	4.8	4	-1	3.6	5	1	4
Florida	7	513	6.4	7	-17	3.2	7	-18	7
California	8	493	4.9	8	-7	3.5	8	-8	8
Alabama	9	466	5.9	9	12	4.0	9	7	9

Note: The U.S. national samples for NAEP and TIMSS include students from both public and private schools.

Table 5b: Actual TIMSS means and predicted means and prediction residual errors for the nine validation states from calibration linking: Science

Jurisdiction	Actual TIMSS Science			(a) TIMSS Science Predicted			(b) Predicted w/ Exclusion Rate Matching		Rank in 2011 NAEP Science
	Rank	Mean	SE	Rank	Residual	SE	Rank	Residual	
U.S. National		525	2.6						
Massachusetts	1	567	5.1	1	-19	3.3	1	-18	1
Minnesota	2	553	4.6	2	-7	3.3	3	-9	2
Colorado	3	542	4.4	3	4	3.9	2	3	3
Indiana	4	533	4.8	5	-6	3.1	5	-6	5
Connecticut	5	532	4.6	4	0	3.5	4	3	4
North Carolina	6	532	6.3	7	-17	3.4	7	-16	7
Florida	7	530	7.3	6	-13	3.5	6	-12	6
California	8	499	4.6	8	0	3.7	8	-5	8
Alabama	9	485	6.2	9	11	3.9	9	5	9

Note: The U.S. national samples for NAEP and TIMSS include students from both public and private schools.

Table 6a: PRESS and MSE values for the predicted means of the nine validation states based on calibration linking: Mathematics

Prediction Approach	**PRESS**	**MSE Prediction**
(a) Calibration Linking	1644	140
(b) Calibration Linking with exclusion rate matching	1403	114

Table 6b: PRESS and MSE values for the predicted means of the nine validation states based on calibration linking: Science

Prediction Approach	**PRESS**	**MSE Prediction**
(a) Calibration Linking	1054	75
(b) Calibration Linking with exclusion rate matching	897	58

Linking Methodology: Statistical Moderation

The following describes the statistical moderation technique applied to establish a link between the 2011 NAEP and the 2011 TIMSS in grade 8 in mathematics and science. In this approach, NAEP results are expressed in the metric of TIMSS. By expressing both assessments in the same metric, estimated the state TIMSS means and percentages of students by TIMSS benchmarks that each state might have obtained had that state actually taken TIMSS. The 2011 NAEP-TIMSS linking using statistical moderation was accomplished in five steps. (Please Note: Step 1 and step 2 correspond to the first stage adjustment, and step 3 corresponds to the second stage adjustment referred to in the highlights report, *U.S. States in a Global Context: Results From the 2011 NAEP-TIMSS Linking Study*, NCES 2013-460.)

Step 1: **Estimating State TIMSS-Equivalent Means from State NAEP Means**

In the discussion below *x* = NAEP and *y* = TIMSS are used in the formulas. In the study by Johnson, Cohen, Chen, Jiang, & Zhang (2005), NAEP was linked to TIMSS using statistical moderation. The same methodology is used in the 2011 NAEP/2011 TIMSS linking study. This means the estimated scores are actually NAEP scores adjusted to have the same mean and standard deviation as TIMSS. That is what it means in *statistical moderation* to say "NAEP is linked to TIMSS." The state mean TIMSS-equivalent $\overline{z}_{1j}$ associated with a NAEP state mean $\overline{x}_j$ is

$$\overline{z}_{1j} = \left(\overline{y} - \frac{\hat{\sigma}_y}{\hat{\sigma}_x} \overline{x} \right) + \left(\frac{\hat{\sigma}_y}{\hat{\sigma}_x} \right) \overline{x}_j \qquad \text{(A1)}$$

$$\hat{A} = \overline{y} - \frac{\hat{\sigma}_y}{\hat{\sigma}_x} \overline{x}$$
$$\hat{B} = \frac{\hat{\sigma}_y}{\hat{\sigma}_x} \qquad \text{(A2)}$$

In equations (A1) and (A2),

- $\hat{A}$ 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. NAEP and U.S. TIMSS results;
- $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are the public school standard deviations NAEP and TIMSS respectively; and
- $\overline{z}_{1j}$ is the mean TIMSS-equivalent of the NAEP mean $\overline{x}_j$ in state *j*.

The error variances in the mean TIMSS-equivalents are

$$\hat{\sigma}^2_{\overline{z}_{1j}} = \hat{B}^2 \hat{\sigma}^2_{\overline{x}_j} + \hat{\sigma}^2_A + 2\left(\overline{x}_j\right)\hat{\sigma}_{AB} + \left(\overline{x}_j\right)^2 \hat{\sigma}^2_B \tag{A3}$$

The square root of equation (A3) is the standard error of linking and forms the basis for the standard errors reported in Tables 15, 16, 19, and 20. According to Johnson et al. (2005), the error variances of the parameters of the linear transformation, $\hat{\sigma}^2_A$, $\hat{\sigma}^2_{AB}$ and $\hat{\sigma}^2_B$, can be approximated by Taylor-series linearization (Wolter, 1985).

$$\begin{aligned}
\hat{\sigma}^2_A &= \hat{B}^2 \hat{\sigma}^2_{\overline{x}} + \hat{\sigma}^2_{\overline{y}} + \overline{x}^2 \hat{B}^2 \left[\frac{\hat{\sigma}^2_{\sigma_y}}{\hat{\sigma}^2_y} + \frac{\hat{\sigma}^2_{\sigma_x}}{\hat{\sigma}^2_x} \right] \\
\hat{\sigma}_{AB} &= -\overline{x}\hat{B}^2 \left[\frac{\hat{\sigma}^2_{\sigma_y}}{\hat{\sigma}^2_y} + \frac{\hat{\sigma}^2_{\sigma_x}}{\hat{\sigma}^2_x} \right] \\
\hat{\sigma}^2_B &= \hat{B}^2 \left[\frac{\hat{\sigma}^2_{\sigma_y}}{\hat{\sigma}^2_y} + \frac{\hat{\sigma}^2_{\sigma_x}}{\hat{\sigma}^2_x} \right]
\end{aligned} \tag{A4}$$

Estimates of the Means and Standard Deviations. The process begins with the analysis of plausible values for both NAEP and TIMSS. In this study, only public school students were included in the analysis of plausible values for both NAEP and TIMSS. In both NAEP and TIMSS, five plausible values were used to represent the student's posterior distribution. Let us label the parameter we are estimating 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 $\overline{p} = \sum_{n=1}^{N} \frac{p_n}{N}$. This formula was used to estimate the means and standard deviations in Table 7 and the linking parameter estimates in Table 11. Table 7 shows the calculations for the parameter estimates of the means and standard deviations.

Table 7: Estimating the Mean and Standard Deviation in U.S. National Samples

	Plausible Value 1	Plausible Value 2	Plausible Value 3	Plausible Value 4	Plausible Value 5	Mean Plausible Value $(\overline{p})$
NAEP mathematics mean	282.78	282.68	282.67	282.77	282.73	282.727
TIMSS mathematics mean	506.17	506.90	507.41	507.20	506.75	506.886
NAEP mathematics SD	36.28	36.30	36.33	36.11	36.23	36.251
TIMSS mathematics SD	75.45	76.34	76.33	75.85	76.22	76.038
NAEP science mean	150.76	150.74	150.77	150.77	150.66	150.741
TIMSS science mean	522.22	521.59	522.31	521.79	523.03	522.188
NAEP science SD	34.44	34.46	34.53	34.53	34.52	34.496
TIMSS science SD	80.95	80.13	79.86	80.28	80.87	80.419

Error variance (sampling). Let us label the error variance due to sampling as *S.* For example, the error variances for the parameter estimates of the means and standard deviations are shown in Table 8. The sampling error in the estimates of the means and standard deviations were obtained 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 deviations 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 8, which are estimates of error variance due to sampling for the mean and standard deviations. This same process was carried out for error variances due to sampling for the linking parameters estimates in Table 12.

Table 8: Sampling Error Variance of the Mean and Standard Deviation $\left(S_{\mu}, S_{\sigma}\right)$

Variance of NAEP mean 2011 mathematics from jackknife	0.0354
Variance of TIMSS mean 2011 mathematics from jackknife	6.6613
Variance of NAEP SD 2011 mathematics from jackknife	0.0218
Variance of TIMSS SD 2011 mathematics from jackknife	2.3423
Variance of NAEP mean 2011 science from jackknife	0.050
Variance of TIMSS mean 2011 science from jackknife	6.034
Variance of NAEP SD 2011 science from jackknife	0.026
Variance of TIMSS SD 2011 science from jackknife	1.770

Error variance (measurement). Let us label the error variance due to measurement as *M.* For example, the error variance for the parameter estimates of the means and standard deviations due to measurement are shown in Table 9. This is estimated by $M = \frac{1+\left(1/N\right)}{N-1}\sum_{n=1}^{N}\left(p_n - \overline{p}\right)^2$. This same process was carried out for error variances due to measurement for the linking parameters estimates in Table 13.

Table 9: Measurement Error Variance of the Mean and Standard Deviation $\left(M_{\mu}, M_{\sigma}\right)$

Variance of NAEP mean 2011 mathematics from plausible values	0.003
Variance of TIMSS mean 2011 mathematics from plausible values	0.273
Variance of NAEP SD 2011 mathematics from plausible values	0.009
Variance of TIMSS SD 2011 mathematics from plausible values	0.177
Variance of NAEP mean 2011 science from plausible values	0.003
Variance of TIMSS mean 2011 science from plausible values	0.368
Variance of NAEP SD 2011 science from plausible values	0.002
Variance of TIMSS SD 2011 science from plausible values	0.268

Error variance (total) of the mean and standard deviation. The total error variance is $T = S + M$ and is shown in Table 10.

Table 10: Total Error Variance of the Mean and Standard Deviation $\left(T_{\mu}, T_{\sigma}\right)$

Variance of NAEP mean 2011 mathematics	0.038
Variance of TIMSS mean 2011 mathematics	6.934
Variance of NAEP SD 2011 mathematics	0.031
Variance of TIMSS SD 2011 mathematics	2.519
Variance of NAEP mean 2011 science	0.053
Variance of TIMSS mean 2011 science	6.402
Variance of NAEP SD 2011 science	0.028
Variance of TIMSS SD 2011 science	2.037

Estimates of the linking parameters A and B. The linking parameters are calculated for each plausible value using equation (A2). The linking parameter estimates are then averaged over the five plausible values as reported in Table 11. Estimates of sampling variance are shown in Table 12. Estimates presented in Table 12 and Table 14 were obtained from equation (A4). Each component of equation (A4) was calculated using procedures described above in the error variance (sampling) and error variance (measurement) section. Estimates presented in Table 14 were obtained as sums of values from Tables 12 and 13.

Table 11: Estimating the Linking Parameters *A* and *B* in the U.S. national samples

	Plausible Value 1	Plausible Value 2	Plausible Value 3	Plausible Value 4	Plausible Value 5	Mean Plausible Value $(\bar{p})$
$\hat{A}$ (mathematics)	–81.963	–87.570	–86.450	–86.669	–88.073	–86.145
$\hat{B}$ (mathematics)	2.080	2.103	2.101	2.100	2.104	2.098
$\hat{A}$ (science)	167.855	171.076	173.627	171.19	170.12	170.77
$\hat{B}$ (science)	2.351	2.325	2.313	2.3	2.3	2.3

Estimates presented in Table 12 and Table 14 were obtained from equation (A4); estimates presented in Table 14 were obtained as sums of values from Table 12 and 13; estimates presented in Tables 13, 15, and 16 were obtained from equation (A1).

Table 12: Sampling Error Variance in *A* and *B* Linking Parameters (S_A, S_B, S_{AB})

Sampling error variance for mathematics in *A*, $(\hat{\sigma}^2_{A(S)})$	155.141
Covariance between *A* and *B* for mathematics, $(\hat{\sigma}_{AB(S)})$	–0.525
Sampling error variance for mathematics in *B*, $(\hat{\sigma}^2_{B(S)})$	0.002
Sampling error variance for science in *A*, $(\hat{\sigma}^2_{A(S)})$	42.805
Covariance between *A* and *B* for science, $(\hat{\sigma}_{AB(S)})$	–0.242
Sampling error variance for science in *B*, $(\hat{\sigma}^2_{B(S)})$	0.002

Error variance (measurement) of the linking parameters A and B. The quantities needed to estimate the error variance in the linking parameters due to measurement error are shown in Table 13 and Table 14. Tables 15 and 16 show standard error estimates for the nine validation states.

Table 13: Measurement Error Variance in *A* and *B* Linking Parameters (M_A, M_B, M_{AB})

Measurement error variance for mathematics in *A*, $(\hat{\sigma}^2_{A(M)})$	13.366
Covariance between *A* and *B* for mathematics, $(\hat{\sigma}_{AB(M)})$	–0.046
Measurement error variance for mathematics in *B*, $(\hat{\sigma}^2_{B(M)})$	0.000
Measurement error variance for science in *A*, $(\hat{\sigma}^2_{A(M)})$	5.725
Covariance between *A* and *B* for science, $(\hat{\sigma}_{AB(M)})$	–0.035
Measurement error variance for science in *B*, $(\hat{\sigma}^2_{B(M)})$	0.000

Table 14: Total Error Variance in *A* and *B* Linking Parameters $\left(T_A, T_B, T_{AB}\right)$

Total error variance for mathematics in *A*, $\left(\hat{\sigma}_A^2\right)$	168.506
Covariance between *A* and *B* for mathematics, $\left(\hat{\sigma}_{AB}\right)$	–0.571
Total error variance for mathematics in *B*, $\left(\hat{\sigma}_B^2\right)$	0.002
Total error variance for science in *A*, $\left(\hat{\sigma}_A^2\right)$	48.531
Covariance between *A* and *B* for science, $\left(\hat{\sigma}_{AB}\right)$	–0.278
Total error variance for science in *B*, $\left(\hat{\sigma}_B^2\right)$	0.002

In Tables 15, 16, 19 and 20, the standard error of the Z-test is based on combining the standard error of the estimate due to linking with the standard error of the actual TIMSS estimate. This is because we are comparing the TIMSS estimate due to linking with the actual TIMSS estimate in the state TIMSS sample. Therefore, the Z-test must incorporate the standard error of the TIMSS estimate due to linking as well as the standard error of the actual TIMSS estimate. This is reflected in the footnote in each table.

Table 15: TIMSS-Equivalents of Nine State NAEP Means in Mathematics

State	TIMSS-Equivalent State Mean	Standard Error Linking	Actual TIMSS Mean	Standard Error TIMSS	Overall Standard Error	z-Test	Significant Difference
Alabama	478	4.0	466	5.9	7.1	1.73	NS
California	486	3.7	493	4.9	6.1	-1.08	NS
Colorado	526	3.5	518	4.9	6.1	1.32	NS
Connecticut	516	3.5	518	4.8	6.0	-0.30	NS
Florida	497	3.2	513	6.4	7.2	-2.32	Significant
Indiana	512	3.4	522	5.1	6.1	-1.60	NS
Massachusetts	540	3.2	561	5.3	6.2	-3.32	Significant
Minnesota	533	3.4	545	4.6	5.7	-2.13	Significant
North Carolina	514	3.4	537	6.8	7.7	-2.95	Significant

Note: Z-test combines the SE due to linking with the actual SE of the TIMSS state mean.

Table 16: TIMSS-Equivalents of Nine State NAEP Means in Science

State	TIMSS-Equivalent State Mean	Standard Error Linking	Actual TIMSS Mean	Standard Error TIMSS	Overall Standard Error	z-Test	Significant Difference
Alabama	497	4.2	485	6.2	7.5	1.57	NS
California	498	4.0	499	4.6	6.1	-0.07	NS
Colorado	545	4.0	542	4.4	5.9	0.54	NS
Connecticut	531	3.7	532	4.6	5.9	-0.04	NS
Florida	517	3.7	530	7.3	8.2	-1.61	NS
Indiana	527	3.3	533	4.8	5.8	-0.94	NS
Massachusetts	547	3.7	567	5.1	6.3	-3.19	Significant
Minnesota	546	3.5	553	4.6	5.8	-1.27	NS
North Carolina	515	3.6	532	6.3	7.2	-2.26	Significant

Note: Z-test combines the SE due to linking with the actual SE of the TIMSS state mean.

Step 2: **Adjusting the State TIMSS-Equivalent Means to Account for Differences in Accommodation Rates between NAEP and TIMSS**

HumRRO conducted an investigation of the relationships between state-level accommodation rates and mean scores and recommended that the state TIMSS-equivalent means be adjusted to account for differences in the accommodation rates among states that predict differences between NAEP and TIMSS exclusion rates. The derivations of specific adjustments are described in a later section of this report, *Adjustments to Predicted State Mean Estimates*, on pages 84 and 85. The following adjustments were used following the HumRRO recommendation.

- For mathematics: $\hat{T}_{Math\ adj}(j) = \hat{T}(j) + (2.65 * (\%\ Acc_j - 9.7))$
  where % Acc_j is the percentage of students in state *j* receiving NAEP accommodations; and 9.7 is the national NAEP accommodation rate for mathematics.

- For science: $\hat{T}_{Science\ adj}(j) = \hat{T}(j) + (2.21 * (\%\ Acc_j - 10.6))$
  where % Acc_j is the percentage of students in state *j* receiving NAEP accommodations; and 10.6 is the national NAEP accommodation rate for science.

The estimated state accommodation rates are shown in Tables 17 and 18.

Table 17: Accommodation Rates in Mathematics

State	Accommodation Rate Mathematics
Alabama	0.04
Alaska	0.14
Arizona	0.09
Arkansas	0.12
California	0.07
Colorado	0.10
Connecticut	0.12
Delaware	0.11
District of Columbia	0.15
DoDEA	0.08
Florida	0.16
Georgia	0.07
Hawaii	0.11
Idaho	0.07
Illinois	0.12
Indiana	0.12
Iowa	0.14
Kansas	0.09
Kentucky	0.08
Louisiana	0.13
Maine	0.14
Maryland	0.07
Massachusetts	0.15
Michigan	0.08
Minnesota	0.09
Mississippi	0.06
Missouri	0.10
Montana	0.09
U.S. National	0.10
National Private	0.05
National Public	0.10
Nebraska	0.09
Nevada	0.09
New Hampshire	0.14
New Jersey	0.14
New Mexico	0.10
New York	0.18
North Carolina	0.12
North Dakota	0.09
Ohio	0.10
Oklahoma	0.04
Oregon	0.11
Pennsylvania	0.13
Rhode Island	0.13
South Carolina	0.08
South Dakota	0.07
Tennessee	0.08
Texas	0.05
Utah	0.08
Vermont	0.15
Virginia	0.09
Washington	0.10
West Virginia	0.09
Wisconsin	0.14
Wyoming	0.11

Table 18: Accommodation Rates in Science

State	Accommodation Rate Science
Alabama	0.04
Alaska	0.16
Arizona	0.09
Arkansas	0.12
California	0.08
Colorado	0.10
Connecticut	0.13
Delaware	0.12
District of Columbia	0.18
DoDEA	0.10
Florida	0.16
Georgia	0.08
Hawaii	0.11
Idaho	0.07
Illinois	0.12
Indiana	0.13
Iowa	0.14
Kansas	0.09
Kentucky	0.08
Louisiana	0.13
Maine	0.14
Maryland	0.11
Massachusetts	0.16
Michigan	0.08
Minnesota	0.08
Mississippi	0.06
Missouri	0.10
Montana	0.09
U.S. National	0.11
National Private	0.05
National Public	0.11
Nebraska	0.12
Nevada	0.11
New Hampshire	0.13
New Jersey	0.17
New Mexico	0.10
New York	0.18
North Carolina	0.12
North Dakota	0.10
Ohio	0.12
Oklahoma	0.10
Oregon	0.10
Pennsylvania	0.15
Rhode Island	0.14
South Carolina	0.09
South Dakota	0.08
Tennessee	0.10
Texas	0.08
Utah	0.09
Vermont	0.14
Virginia	0.10
Washington	0.10
West Virginia	0.09
Wisconsin	0.14
Wyoming	0.11

The TIMSS-equivalents of the nine validation state NAEP means with adjustments for accommodations are contained in the Tables 19 and 20.

Table 19: TIMSS-Equivalents of State NAEP Means with Adjustments for Accommodations in Mathematics

State	Predicted TIMSS-State Mean	Standard Error Linking	Actual TIMSS State Mean	Standard Error State TIMSS	Overall Standard Error	z-Test	Significant Difference
Alabama	462	4.0	466	5.9	7.1	-0.53	NS
California	480	3.7	493	4.9	6.1	-2.04	Significant
Colorado	527	3.5	518	4.9	6.1	1.45	NS
Connecticut	523	3.5	518	4.8	6.0	0.85	NS
Florida	514	3.2	513	6.4	7.2	0.06	NS
Indiana	518	3.4	522	5.1	6.1	-0.53	NS
Massachusetts	554	3.2	561	5.3	6.2	-1.05	NS
Minnesota	530	3.4	545	4.6	5.7	-2.60	Significant
North Carolina	521	3.4	537	6.8	7.7	-2.02	Significant

Note: Z-test combines the SE due to linking with the actual SE of the TIMSS state mean.

Table 20: TIMSS-Equivalents of State NAEP Means with Adjustments for Accommodations in Science

State	Predicted TIMSS State Mean	Standard Error Linking	Actual TIMSS State Mean	Standard Error State TIMSS	Overall Standard Error	z-Test	Significant Difference
Alabama	483	4.2	485	6.2	7.5	-0.34	NS
California	492	4.0	499	4.6	6.1	-1.09	NS
Colorado	544	4.0	542	4.4	5.9	0.43	NS
Connecticut	536	3.7	532	4.6	5.9	0.69	NS
Florida	529	3.7	530	7.3	8.2	-0.08	NS
Indiana	532	3.3	533	4.8	5.8	-0.06	NS
Massachusetts	558	3.7	567	5.1	6.3	-1.31	NS
Minnesota	541	3.5	553	4.6	5.8	-2.07	Significant
North Carolina	519	3.6	532	6.3	7.2	-1.80	NS

Note: Z-test combines the SE due to linking with the actual SE of the TIMSS state mean.

Step 3: **Predicting State TIMSS Means from Adjusted TIMSS-Equivalents of State NAEP Means**

In the sections above the goal was to link or rescale NAEP to have the same scale as TIMSS. This allows us to find the NAEP score on the NAEP scale that is the TIMSS-equivalent of the

TIMSS international benchmarks Low, Intermediate, High, and Advanced. A second goal of the study is the estimated state performance on TIMSS based on NAEP performance in the 43 states in which TIMSS was not administered at the state level. We can do that in this study by taking advantage of the correlation between NAEP and TIMSS estimated from the nine validation states. The prediction of the state TIMSS means from the state NAEP means can be accomplished through statistical projection.

$$\overline{z}_{2j} = \left(\overline{\overline{y}} - \hat{\rho} \frac{\hat{\sigma}_{\overline{y}}}{\hat{\sigma}_{\overline{z}_1}} \overline{\overline{z}}_1 \right) + \left(\hat{\rho} \frac{\hat{\sigma}_{\overline{y}}}{\hat{\sigma}_{\overline{z}_{\cdot 1}}} \right) \overline{z}_{1j} \qquad \text{(A5)}$$

With intercept and slope regression parameters

$$\hat{\alpha} = \overline{\overline{y}} - \hat{\rho} \frac{\hat{\sigma}_{\overline{y}}}{\hat{\sigma}_{\overline{z}_1}} \overline{\overline{z}}_1$$

$$\hat{\beta} = \hat{\rho} \frac{\hat{\sigma}_{\overline{y}}}{\hat{\sigma}_{\overline{z}_1}} \qquad \text{(A6)}$$

The quantities in equation (A5) are defined as follows:

- $\overline{z}_{2j}$ is the *predicted* state mean TIMSS-equivalent for a given $\overline{z}_{1j}$;
- $\overline{\overline{z}}_1$ is the weighted mean of the adjusted TIMSS-equivalent means (from step 2) among the nine validation states (weighted by the effective sample sizes in each state);
- $\overline{z}_{1j}$ is the adjusted state mean TIMSS-equivalent (from step 2) obtained for each of the validation states;
- $\hat{\sigma}_{\overline{z}_1}$ is the weighted standard deviation of the adjusted state means of TIMSS-equivalents in the nine validation states;
- $\overline{\overline{y}}$ is the weighted mean of the actual TIMSS means among the nine validation states;
- $\hat{\sigma}_{\overline{y}}$ is the weighted standard deviation of the state means of actual TIMSS among the nine validation states; and
- $\hat{\rho}$ is the weighted correlation between the state's mean TIMSS-equivalents $\overline{z}_{1j}$ and actual TIMSS state means $\overline{y}_j$ in the nine validation states.

The error variance in the projection is found by

$$\hat{\sigma}^2_{\overline{z}_{2j}} = \hat{\beta}^2 \hat{\sigma}^2_{\overline{z}_{1j}} + \hat{\sigma}^2_{\hat{\alpha}} + 2\left(\overline{z}_{1j}\right)\hat{\sigma}_{\hat{\alpha},\hat{\beta}} + \left(\overline{z}_{1j}\right)^2 \hat{\sigma}^2_{\hat{\beta}} \cdot \tag{A7}$$

The square root of equation (A7) is the standard error of projection that is presented in Tables 22 and 24.

In equation (A7) the projection error variance components are as follows

- $\hat{\beta}^2$ times the linking error variance $\hat{\sigma}^2_{\overline{z}_{1j}}$ in the TIMSS-equivalents, and
- the prediction error variance (how accurate the α and β were estimated) $\hat{\sigma}^2_{\hat{\alpha}} + 2\left(\overline{z}_{1j}\right)\hat{\sigma}_{\hat{\alpha},\hat{\beta}} + \left(\overline{z}_{1j}\right)^2 \hat{\sigma}^2_{\hat{\beta}} \cdot$

The variances and co-variances of α and β in equation (A7) are

$$\begin{aligned}\hat{\sigma}^2_{\hat{\alpha}} \approx\ & \hat{\beta}^2 Var(\overline{\overline{z}}_1) + \hat{\beta}^2 \frac{\overline{\overline{z}}_1^2}{\hat{\sigma}^2_{\overline{z}_1}} Var(\hat{\sigma}_{\overline{z}_1}) + Var(\overline{\overline{y}}) + \hat{\beta}^2 \frac{\overline{\overline{z}}_1^2}{\hat{\sigma}^2_{\overline{y}}} Var(\hat{\sigma}_{\overline{y}}) + \hat{\beta}^2 \frac{\overline{\overline{z}}_1^2}{\hat{\rho}^2} Var(\hat{\rho}) \\ & -2\hat{\beta} Cov(\overline{\overline{z}}_1, \overline{\overline{y}}) - 2\hat{\beta}^2 \frac{\overline{\overline{z}}_1^2}{\hat{\sigma}_{\overline{z}_1}\hat{\sigma}_{\overline{y}}} Cov(\hat{\sigma}_{\overline{z}_1}, \hat{\sigma}_{\overline{y}}) - 2\hat{\beta}^2 \frac{\overline{\overline{z}}_1^2}{\hat{\rho}\hat{\sigma}_{\overline{z}_1}} Cov(\hat{\sigma}_{\overline{z}_1}, \hat{\rho}) \\ & +2\hat{\beta}^2 \frac{\overline{\overline{z}}_1^2}{\hat{\rho}\hat{\sigma}_{\overline{y}}} Cov(\hat{\sigma}_{\overline{y}}, \hat{\rho})\end{aligned} \tag{A8}$$

$$\begin{aligned}\hat{\sigma}^2_{\hat{\beta}} \approx\ & \hat{\beta}^2 \frac{1}{\hat{\sigma}^2_{\overline{z}_1}} Var(\hat{\sigma}_{\overline{z}_1}) + \hat{\beta}^2 \frac{1}{\hat{\sigma}^2_{\overline{y}}} Var(\hat{\sigma}_{\overline{y}}) + \hat{\beta}^2 \frac{1}{\hat{\rho}^2} Var(\hat{\rho}) \\ & -2\hat{\beta}^2 \frac{1}{\hat{\sigma}_{\overline{z}_1}\hat{\sigma}_{\overline{y}}} Cov(\hat{\sigma}_{\overline{z}_1}, \hat{\sigma}_{\overline{y}}) - 2\hat{\beta}^2 \frac{1}{\hat{\rho}\hat{\sigma}_{\overline{z}_1}} Cov(\hat{\sigma}_{\overline{z}_1}, \hat{\rho}) \\ & +2\hat{\beta}^2 \frac{1}{\hat{\rho}\hat{\sigma}_{\overline{y}}} Cov(\hat{\sigma}_{\overline{y}}, \hat{\rho})\end{aligned} \tag{A9}$$

$$\hat{\sigma}_{\hat{\alpha},\hat{\beta}} \approx -\hat{\beta}^2 \frac{\overline{\overline{z}}_1}{\hat{\sigma}^2_{\overline{z}_1}} Var(\hat{\sigma}_{\overline{z}_1}) - \hat{\beta}^2 \frac{\overline{\overline{z}}_1}{\hat{\sigma}^2_{\overline{y}}} Var(\hat{\sigma}_{\overline{y}}) - \hat{\beta}^2 \frac{\overline{\overline{z}}_1}{\hat{\rho}^2} Var(\hat{\rho})$$
$$+2\hat{\beta}^2 \frac{\overline{\overline{z}}_1}{\hat{\sigma}_{\overline{z}_1}\hat{\sigma}_{\overline{y}}} Cov(\hat{\sigma}_{\overline{z}_1}, \hat{\sigma}_{\overline{y}}) + 2\hat{\beta}^2 \frac{\overline{\overline{z}}_1}{\hat{\rho}\hat{\sigma}_{\overline{z}_1}} Cov(\hat{\sigma}_{\overline{z}_1}, \hat{\rho}) \qquad \text{(A10)}$$
$$-2\hat{\beta}^2 \frac{\overline{\overline{z}}_1}{\hat{\rho}\hat{\sigma}_{\overline{y}}} Cov(\hat{\sigma}_{\overline{y}}, \hat{\rho})$$

The components of equations (A8) to (A10) can be estimated as follows:

$$Var(\overline{\overline{z}}_1) = \frac{\hat{\sigma}^2_{\overline{z}_1}}{n}$$

$$Var(\overline{\overline{y}}) = \frac{\hat{\sigma}^2_{\overline{y}}}{n}$$

$$Var(\hat{\sigma}_{\overline{z}_1}) = \frac{\hat{\sigma}^2_{\overline{z}_1}}{2(n-1)}$$

$$Var(\hat{\sigma}_{\overline{y}}) = \frac{\hat{\sigma}^2_{\overline{y}}}{2(n-1)}$$

$$Var\left(\hat{\sigma}^2_{\overline{z}_1}\right) \approx 4\hat{\sigma}^2_{\overline{z}_1} Var\left(\hat{\sigma}_{\overline{z}_1}\right)$$

$$Var\left(\hat{\sigma}^2_{\overline{y}}\right) \approx 4\hat{\sigma}^2_{\overline{y}} Var\left(\hat{\sigma}_{\overline{y}}\right)$$

$$Var\left(\hat{\rho}\right) \approx \left(1-\hat{\rho}^2\right)^2 \left\{ \frac{1}{n-1} + \frac{11\hat{\rho}^2}{2(n-1)^2} + \frac{-24\hat{\rho}^2 + 75\hat{\rho}^4}{1 \quad (n-61)^3} \right\}$$

$$Var\left(\hat{\rho}^2\right) = 4\hat{\rho}^2 Var\left(\hat{\rho}\right)$$

$$Cov(\overline{\overline{z}}_1, \overline{\overline{y}}) = \hat{\rho}\sqrt{Var(\overline{\overline{z}}_1)Var(\overline{\overline{y}})}$$

$$Cov\left(\hat{\sigma}_{\overline{z}_1}, \hat{\sigma}_{\overline{y}}\right) \approx \hat{\rho}^2\sqrt{Var\left(\hat{\sigma}_{\overline{z}_1}\right)Var\left(\hat{\sigma}_{\overline{y}}\right)}$$

$$Cov\left(\hat{\sigma}^2_{\overline{z}_1}, \hat{\sigma}^2_{\overline{y}}\right) \approx \hat{\rho}^2\sqrt{Var\left(\hat{\sigma}^2_{\overline{z}_1}\right)Var\left(\hat{\sigma}^2_{\overline{y}}\right)}$$

$$Cov\left(\hat{\rho}, \hat{\sigma}_{\overline{z}_1}\right) \approx \hat{\rho}_{\hat{\rho}^2, \hat{\sigma}^2_{\overline{z}_1}}\sqrt{Var\left(\hat{\rho}\right)Var\left(\hat{\sigma}_{\overline{z}_1}\right)}$$

$$Cov\left(\hat{\rho}^2, \hat{\sigma}^2_{\overline{z}_1}\right) \approx \hat{\rho}_{\hat{\rho}^2, \hat{\sigma}^2_{\overline{z}_1}}\sqrt{Var\left(\hat{\rho}^2\right)Var\left(\hat{\sigma}^2_{\overline{z}_1}\right)}$$

$$\hat{\rho}_{\hat{\rho}^2, \hat{\sigma}^2_{\overline{z}_1}} \approx \frac{\left(\hat{\rho}^2\hat{\sigma}^2_{\overline{z}_1} + \frac{\left(1-\hat{\rho}^2\right)}{n-1}\hat{\sigma}^2_{\overline{z}_1} - \left(Var\left(\hat{\rho}\right) + \hat{\rho}^2\right)\left(Var\left(\hat{\sigma}_{\overline{z}_1}\right) + \hat{\sigma}^2_{\overline{z}_1}\right)\right)}{\sqrt{Var\left(\hat{\sigma}^2_{\overline{z}_1}\right)Var\left(\hat{\rho}^2\right)}}$$

$$Cov\left(\hat{\rho}, \hat{\sigma}_{\overline{y}}\right) \approx \hat{\rho}_{\hat{\rho}^2, \hat{\sigma}^2_{\overline{y}}}\sqrt{Var\left(\hat{\rho}\right)Var\left(\hat{\sigma}_{\overline{y}}\right)}$$

$$Cov\left(\hat{\rho}, \hat{\sigma}^2_{\overline{y}}\right) \approx \hat{\rho}_{\hat{\rho}^2, \hat{\sigma}^2_{\overline{y}}}\sqrt{Var\left(\hat{\rho}\right)Var\left(\hat{\sigma}^2_{\overline{y}}\right)}$$

$$\hat{\rho}_{\hat{\rho}^2,\hat{\sigma}^2_{\bar{y}}} \approx \frac{\left(\hat{\rho}^2\hat{\sigma}^2_{\bar{y}} + \frac{\left(1-\hat{\rho}^2\right)}{n-1}\hat{\sigma}^2_{\bar{y}} - \left(Var\left(\hat{\rho}\right)+\hat{\rho}^2\right)\left(Var\left(\hat{\sigma}_{\bar{y}}\right)+\hat{\sigma}^2_{\bar{y}}\right)\right)}{\sqrt{Var\left(\hat{\sigma}^2_{\bar{y}}\right)Var\left(\hat{\rho}^2\right)}}$$

Weighted correlations between the TIMSS-equivalent means and the actual TIMSS means for the nine validation states were calculated with and without accommodation adjustments. Without accommodation adjustments, the weighted correlations were .92 and .93 for mathematics and science, respectively. After the accommodation adjustments were applied to the nine states, the weighted correlations were .94 and .97 for mathematics and science, respectively. In both cases the weighted correlation between TIMSS-equivalent means and actual TIMSS means were improved by the adjustment for accommodations. Therefore, the accommodation adjustments in both mathematics and science were used. Below, the projections are conducted for the nine validation states with the accommodation adjustments. Tables 21 and 23 show the projection parameter estimates and Tables 22 and 24 show the resulting state mean estimates and standard errors for mathematics and science respectively.

The standard errors in Table 22 contained two components, standard error of linking and standard error of prediction. The error variance of linking was given in equation (A3), and the error variance of prediction was given in equation (A7), which estimates the degree of uncertainty in the prediction equation. Note that in the section *Evaluations of the Methodologies* another source of error—model error—was discussed. Model error is a valuable criterion in quantifying the discrepancies between actual and predicted values. However, in reporting the predicted TIMSS scores for the 43 states that did not participate in TIMSS at the state level, the model error in standard error calculation and hypothesis testing was not included. This is because model error variance reflected estimation bias as well as variability/standard error, and the data at the state level necessary to evaluate and account for bias was limited for the validation states.

Table 21: Projection Parameters for Mathematics Means

Correlation	Parameter Estimates	
	Alpha	*Beta*
0.94	32.1584	0.9457
Variance-Covariance		
	Alpha	*Beta*
Alpha	15.1720	-0.0294
Beta	-0.0294	0.0001

Table 22: Projection for Mathematics with Accommodation Adjustments

State	Projection	Standard Error Projection	Actual TIMSS	Standard Error TIMSS	Overall Standard Error	z-Test	Significant Difference
Alabama	469	3.8	466	5.9	7.0	0.46	NS
California	486	3.5	493	4.9	6.0	-1.06	NS
Colorado	530	3.4	518	4.9	5.9	2.07	Significant
Connecticut	526	3.3	518	4.8	5.9	1.51	NS
Florida	518	3.0	513	6.4	7.1	0.66	NS
Indiana	522	3.2	522	5.1	6.0	0.12	NS
Massachusetts	556	3.1	561	5.3	6.1	-0.72	NS
Minnesota	533	3.2	545	4.6	5.6	-2.05	Significant
North Carolina	525	3.2	537	6.8	7.6	-1.53	NS

Note: Z-test combines the SE due to projection with the actual SE of the TIMSS estimate. The standard error of projection is the square root of equation A7.

Table 23: Projection Parameters for Science Means

Correlation		
	Alpha	*Beta*
0.97	20.3460	0.9680
Variance-Covariance		
	Alpha	*Beta*
Alpha	7.9064	-0.0150
Beta	-0.0150	0.0000

Table 24: Projection for Science with Accommodation Adjustments

State	Projection	Standard Error Projection	Actual TIMSS	Standard Error TIMSS	Overall Standard Error	z-Test	Significant Difference
Alabama	488	4.1	485	6.2	7.4	0.31	NS
California	496	3.9	499	4.6	6.0	-0.34	NS
Colorado	547	3.8	542	4.4	5.8	0.94	NS
Connecticut	539	3.6	532	4.6	5.8	1.26	NS
Florida	533	3.6	530	7.3	8.1	0.34	NS
Indiana	536	3.2	533	4.8	5.7	0.52	NS
Massachusetts	561	3.6	567	5.1	6.2	-0.93	NS
Minnesota	544	3.4	553	4.6	5.8	-1.57	NS
North Carolina	522	3.5	532	6.3	7.2	-1.29	NS

Note: Z-test combines the SE due to projection with the actual SE of the TIMSS estimate. The standard error of projection is the square root of equation A7.

Step 4: **Estimating the Percentages at and above International Benchmarks in the State TIMSS-Equivalent Distribution (after adjustments for accommodations)**

The distribution of z_{1j} in each state (after adjustments for accommodations) can be determined from equation (A1) by substituting z_{1j} for $\overline{z}_{1j}$ and x_j for $\overline{x}_j$. Once the distribution of z_{1j} is determined, we can estimate the proportion above various cut-scores on z_{1j}. For example, if z_{1j} scores are TIMSS-equivalents of State-NAEP scores, then $1-\hat{p}_{1j}$ is the proportion of students in the state we estimate would be above the international benchmarks on TIMSS-equivalents in each state. The quantity $1-\hat{p}_{1j}$ can be estimated via a normal approximation.

$$
\begin{aligned}
1-\hat{p}_{1j} &= \Pr\left(z_{1j} \geq z_{benchmark}\right) \\
&= \int_{-\infty}^{\infty} \Pr\left(z_{1j} \geq z_{benchmark} \mid x_j\right) f\left(x_j \mid \overline{x}_j, \hat{\sigma}^2_{X_j}\right) dx_j \qquad \text{(A11)} \\
&= \int_{z_{benchmark}}^{\infty} f\left(x_j \mid A + B\overline{x}_j, B^2\hat{\sigma}^2_{X_j}\right) dx_j
\end{aligned}
$$

We can define $h(z_{benchmark}, \overline{z}_{1j}, \hat{\sigma}_{z_{1j}}) = \int_{z_{benchmark}}^{\infty} f\left(x_j \mid A + B\overline{x}_j, B^2\hat{\sigma}^2_{X_j}\right) dx_j$. The linking error variance in z_{1j} will be propagated to $1-\hat{p}_{1j}$. Using Taylor series approximation, the error variance of $1-\hat{p}_{1j}$ due to linking is

$$
\begin{aligned}
\sigma^2_{L(1-p_{1j})} &= Var(h(z_{benchmark}, \overline{z}_{1j}, \hat{\sigma}_{z_{1j}})) \\
&\approx \left(\frac{\exp\left(-(z_{benchmark} - \overline{z}_{1j})^2 / (2\hat{\sigma}^2_{z_{1j}})\right)}{\sqrt{2\pi}\hat{\sigma}_{z_{1j}}}\right)^2 Var\left(z_{1j}\right) \\
&+ \left(\frac{\exp\left(-(z_{benchmark} - \overline{z}_{1j})^2 / (2\hat{\sigma}^2_{z_{1j}})\right)}{\sqrt{2\pi}\hat{\sigma}_{z_{1j}}}\right)^2 Var\left(\overline{z}_{1j}\right) \qquad \text{(A12)} \\
&+ \left(\left(\frac{z_{benchmark} - \overline{z}_{1j}}{\hat{\sigma}_{z_{1j}}}\right)\frac{\exp\left(-(z_{benchmark} - \overline{z}_{1j})^2 / (2\hat{\sigma}^2_{z_{1j}})\right)}{\sqrt{2\pi}\hat{\sigma}_{z_{1j}}}\right)^2 Var\left(\hat{\sigma}_{z_{1j}}\right)
\end{aligned}
$$

In the above equation

- z_{1j} is the TIMSS-equivalent of the NAEP score x_j;
- $Var\left(z_{1j}\right)$ is the linking error variance in z_1 obtained by $Var\left(z_{1j}\right) = \hat{B}^2\hat{\sigma}^2_{x_j} + \hat{\sigma}^2_A + 2\left(x_j\right)\hat{\sigma}_{AB} + \left(x_j\right)^2\hat{\sigma}^2_B$;
- $Var\left(\overline{z}_{1j}\right)$ is the error variance in the mean of z_{1j}; and
- $Var\left(\hat{\sigma}_{z_{1j}}\right)$ is the error variance in the standard deviation of z_{1j}.

Step 5: **Predicting the Percentages at and Above International Benchmarks**

Predicting the percentages at and above international benchmarks in the projected distribution $1-p_{2j}$ uses equations (A5), (A6), and (A7) with the following substitutions

- $1-p_{1j}$ (the percentages at and above in the TIMSS-equivalent distribution) is substituted for $\overline{z}_{1j}$;
- $1-p_{2j}$ (the predicted percentages at and above TIMSS international benchmarks) is substituted for $\overline{z}_{2j}$;
- the mean of $1-p_j$ (the actual percentages at and above) is substituted for $\overline{\overline{y}}$; and
- the mean of $1-p_{1j}$ is substituted for $\overline{\overline{z}}_1$.

The parameter estimates needed to conduct the projections for each of the international benchmarks (step 5) and the resulting distributions for the nine validation states are contained in Tables 25 through 40 below.

Table 25: Projection Parameters for Low International Benchmark in Mathematics

Correlation		
	Alpha	*Beta*
0.90	17.4697	0.8063
Variance-Covariance		
	Alpha	*Beta*
Alpha	0.6641	-0.0071
Beta	-0.0071	0.0001

Table 26: Predicted Estimates of TIMSS Percentages for Low Benchmark with Adjustments for Accommodations in Grade 8 Math

State	Projection	Standard Error Projection	Actual TIMSS	Standard Error TIMSS	Overall Standard Error	z-Test	Significant Difference
Alabama	82	1.8	79	2.2	2.8	1.08	NS
California	85	1.5	87	1.7	2.3	-1.22	NS
Colorado	94	0.8	93	1.1	1.3	0.66	NS
Connecticut	94	0.9	91	1.4	1.7	2.00	Significant
Florida	93	0.9	94	1.3	1.6	-0.35	NS
Indiana	95	1.1	95	1.0	1.5	-0.29	NS
Massachusetts	97	0.5	98	0.3	0.6	-1.51	NS
Minnesota	95	0.7	97	0.7	0.9	-2.37	Significant
North Carolina	94	0.9	95	1.3	1.6	-0.94	NS

Note: Z-test combines the SE of the projected percentages and the SE of the actual TIMSS percentages.

Table 27: Projection Parameters for Intermediate International Benchmark in Mathematics

Correlation	Parameter Estimates	
	Alpha	*Beta*
0.92	10.7261	0.8567
Variance-Covariance		
	Alpha	*Beta*
Alpha	0.3554	-0.0049
Beta	-0.0049	0.0001

Table 28: Predicted Estimates of TIMSS Percentages for Intermediate Benchmark with Adjustments for Accommodations in Grade 8 Math

State	Projection	Standard Error Projection	Actual TIMSS	Standard Error TIMSS	Overall Standard Error	z-Test	Significant Difference
Alabama	48	2.7	46	3.1	4.1	0.47	NS
California	56	2.2	59	2.8	3.5	-0.95	NS
Colorado	75	2.0	71	2.5	3.2	1.45	NS
Connecticut	74	2.1	69	2.5	3.3	1.47	NS
Florida	71	2.2	68	3.3	4.0	0.80	NS
Indiana	74	2.1	74	2.3	3.1	-0.13	NS
Massachusetts	85	1.7	88	1.4	2.2	-1.55	NS
Minnesota	77	1.8	83	1.9	2.6	-2.27	Significant
North Carolina	73	2.0	78	2.5	3.2	-1.39	NS

Note: Z-test combines the SE of the projected percentages and the SE of the actual TIMSS percentages.

Table 29: Projection Parameters for High International Benchmark in Mathematics

Correlation		
	Alpha	*Beta*
0.94	2.1544	1.0356
Variance-Covariance		
	Alpha	*Beta*
Alpha	0.0888	-0.0024
Beta	-0.0024	0.0001

Table 30: Predicted Estimates of TIMSS Percentages for High Benchmark with Adjustments for Accommodations in Grade 8 Math

State	Projection	Standard Error Projection	Actual TIMSS	Standard Error TIMSS	Overall Standard Error	z-Test	Significant Difference
Alabama	15	2.3	15	2.5	3.4	-0.05	NS
California	23	2.2	24	2.5	3.3	-0.49	NS
Colorado	41	2.8	35	2.7	3.9	1.59	NS
Connecticut	39	2.7	37	2.9	4.0	0.64	NS
Florida	34	2.7	31	3.2	4.1	0.78	NS
Indiana	36	2.2	35	3.3	4.0	0.05	NS
Massachusetts	56	2.7	57	3.2	4.2	-0.25	NS
Minnesota	43	2.8	49	2.8	4.0	-1.55	NS
North Carolina	39	2.5	44	3.6	4.4	-1.28	NS

Note: Z-test combines the SE of the projected percentages and the SE of the actual TIMSS percentages.

Table 31: Projection Parameters for Advanced International Benchmark in Mathematics

Correlation	Parameter Estimates	
	Alpha	*Beta*
0.93	0.4132	1.1453
Variance-Covariance		
	Alpha	*Beta*
Alpha	0.0087	-0.0009
Beta	-0.0009	0.0001

Table 32: Predicted Estimates of TIMSS Percentages for Advanced Benchmark with Adjustments for Accommodations in Grade 8 Math

State	Projection	Standard Error Projection	Actual TIMSS	Standard Error TIMSS	Overall Standard Error	z-Test	Significant Difference
Alabama	2	0.8	2	0.8	1.1	-0.01	NS
California	5	1.0	5	0.9	1.4	0.13	NS
Colorado	11	1.8	8	1.1	2.1	1.72	NS
Connecticut	10	1.5	10	1.3	2.0	-0.06	NS
Florida	8	1.3	8	1.6	2.0	-0.04	NS
Indiana	7	0.9	7	1.2	1.5	0.22	NS
Massachusetts	19	2.0	19	3.0	3.6	-0.06	NS
Minnesota	12	1.9	13	2.3	3.0	-0.47	NS
North Carolina	10	1.4	14	2.6	3.0	-1.24	NS

Note: Z-test combines the SE of the projected percentages and the SE of the actual TIMSS percentages.

Table 33: Projection Parameters for Low International Benchmark in Science

Correlation	Parameter Estimates	
	Alpha	*Beta*
0.92	18.2179	0.7977
Variance-Covariance		
	Alpha	*Beta*
Alpha	0.4500	-0.0048
Beta	-0.0048	0.0001

Table 34: Predicted Estimates of TIMSS Percentages for Low Benchmark with Adjustments for Accommodations in Grade 8 Science

State	Projection	Standard Error Projection	Actual TIMSS	Standard Error TIMSS	Overall Standard Error	z-Test	Significant Difference
Alabama	85	1.5	83	1.9	2.4	0.86	NS
California	86	1.5	88	1.6	2.2	-0.59	NS
Colorado	96	0.6	96	0.7	0.9	-0.64	NS
Connecticut	95	0.8	92	1.3	1.5	1.69	NS
Florida	94	0.7	93	1.5	1.7	0.11	NS
Indiana	95	1.0	95	0.9	1.4	-0.09	NS
Massachusetts	96	0.6	96	0.7	0.9	-0.37	NS
Minnesota	96	0.5	98	0.7	0.9	-2.37	Significant
North Carolina	93	1.0	94	1.4	1.7	-0.94	NS

Note: Z-test combines the SE of the projected percentages and the SE of the actual TIMSS percentages.

Table 35: Projection Parameters for Intermediate International Benchmark in Science

Correlation	Parameter Estimates	
	Alpha	*Beta*
0.95	12.1405	0.8437
Variance-Covariance		
	Alpha	*Beta*
Alpha	0.2030	-0.0027
Beta	-0.0027	0.0000

Table 36: Predicted Estimates of TIMSS Percentages for Intermediate Benchmark with Adjustments for Accommodations in Grade 8 Science

State	Projection	Standard Error Projection	Actual TIMSS	Standard Error TIMSS	Overall Standard Error	z-Test	Significant Difference
Alabama	58	2.5	56	3.5	4.3	0.30	NS
California	61	2.2	62	2.5	3.4	-0.38	NS
Colorado	81	1.9	80	2.0	2.7	0.64	NS
Connecticut	78	1.9	74	2.0	2.8	1.32	NS
Florida	75	1.9	74	3.6	4.0	0.39	NS
Indiana	78	2.1	78	2.1	3.0	0.07	NS
Massachusetts	84	1.7	87	1.5	2.3	-1.30	NS
Minnesota	81	1.7	85	2.0	2.6	-1.70	NS
North Carolina	72	2.1	75	3.0	3.6	-0.78	NS

Note: Z-test combines the SE of the projected percentages and the SE of the actual TIMSS percentages.

Table 37: Projection Parameters for High International Benchmark in Science

Correlation	Parameter Estimates	
	Alpha	*Beta*
0.97	1.4500	1.0586
Variance-Covariance		
	Alpha	*Beta*
Alpha	0.0550	-0.0013
Beta	-0.0013	0.0000

Table 38: Predicted Estimates of TIMSS Percentages for High Benchmark with Adjustments for Accommodations in Grade 8 Science

State	Projection	Standard Error Projection	Actual TIMSS	Standard Error TIMSS	Overall Standard Error	z-Test	Significant Difference
Alabama	23	2.7	24	2.7	3.8	-0.11	NS
California	28	2.5	28	1.9	3.2	0.06	NS
Colorado	51	3.2	48	2.6	4.1	0.84	NS
Connecticut	47	2.8	45	2.5	3.8	0.47	NS
Florida	44	2.7	42	3.5	4.4	0.48	NS
Indiana	45	2.7	43	2.9	3.9	0.30	NS
Massachusetts	59	2.8	61	2.8	4.0	-0.65	NS
Minnesota	49	2.9	54	2.6	3.9	-1.10	NS
North Carolina	38	2.6	42	3.2	4.2	-1.04	NS

Note: Z-test combines the SE of the projected percentages and the SE of the actual TIMSS percentages.

Table 39: Projection Parameters for Advanced International Benchmark in Science

Correlation	Parameter Estimates	
	Alpha	*Beta*
0.96	-0.7354	1.1930
Variance-Covariance		
	Alpha	*Beta*
Alpha	0.0088	-0.0007
Beta	-0.0007	0.0001

Table 40: Predicted Estimates of TIMSS Percentages for Advanced Benchmark with Adjustments for Accommodations in Grade 8 Science

State	Projection	Standard Error Projection	Actual TIMSS	Standard Error TIMSS	Overall Standard Error	z-Test	Significant Difference
Alabama	4	1.5	5	1.0	1.8	-0.32	NS
California	7	1.5	6	0.7	1.7	0.54	NS
Colorado	16	2.5	14	1.6	3.0	0.67	NS
Connecticut	15	2.0	14	1.5	2.6	0.19	NS
Florida	13	1.8	13	2.0	2.6	-0.03	NS
Indiana	12	1.7	10	1.4	2.2	0.75	NS
Massachusetts	23	2.4	24	2.6	3.5	-0.32	NS
Minnesota	15	2.3	16	1.9	3.0	-0.54	NS
North Carolina	10	1.6	12	2.2	2.7	-1.01	NS

Note: Z-test combines the SE of the projected percentages and the SE of the actual TIMSS percentages.

Evaluations of the Methodologies

The following section describes the key findings from HumRRO's evaluation and summarizes the evidence underlying those findings. In the following descriptions, the methodologies are referred to **CAL** for the joint calibration method, **PRO** for the statistical projection method, and **MOD** for statistical moderation.

Evaluation Design

Two stages were included in the plan for evaluating the results of the linkage study. The first stage involved applying each of the linkages to state NAEP samples for the nine validation states participating in TIMSS and comparing the resulting estimates to corresponding estimates generated from the operational TIMSS state samples. HumRRO reviewed NAEP and TIMSS reports and concluded that the statistics most likely to be used in reporting results from the linkage were scale score means and the percentage of students at or above each of the TIMSS benchmarks. HumRRO also examined differences in the estimated TIMSS scale score standard deviations for each validation state, providing a general comparison of differences in the estimated scale score distributions throughout the score range. In addition to comparing statistics for each state sample as a whole, they also examined differences in linkage estimates for subgroups defined by gender and, where sample size permitted, race/ethnicity.

The second stage of the evaluation involved investigation of the extent to which key differences between the two assessments affected the linkages or threatened the validity of key interpretations. Prior to analyzing any data, HumRRO conducted discussions with key members of the Quality Assurance Technical Panel (QATP)[6] to identify differences between the two assessments that might plausibly affect the scale score linkages. Figure 4 lists key differences identified in these discussions. Some differences, such as differences in accommodation and exclusion rates, could be readily quantified so that state-level differences could be related to state-level differences in the linkages. Others, such as the impact of the difference in testing windows, could not be investigated directly from the available data. Note that the braided samples did provide estimates from each of the two assessments during each testing window, but the braided samples were too small to support separate analyses by state and also testing window differences were confounded with other differences in test administration procedures.

[6] The QATP comprises nine nationally and internationally recognized experts in various aspects of assessment who work with HumRRO to design and implement special quality assurance studies. Four panelists, in particular, provided ongoing advice on the NAEP-TIMSS linkage: Kadriye Ercikan, Mark Reckase, William Schafer, and Richard Wolfe.

Assessment Process	Differences in…
Content	• Content coverage • Slight differences in item format • Test administration time
Sampling	• Sampling method • Sample size • Minimum acceptable participation rate
Administration	• Administration timing (time of year)
Inclusion and accommodation	• Accommodation policy • Exclusion policy
Analysis and scaling	• Conditioning model • Treatment of not-reached items • Establishing trend
Reporting	• Benchmarks • Scale (Score range, mean, SD)

Figure 4. Key differences between the NAEP and TIMSS assessments.

Primary Findings and Conclusions – Stage 1

Tables 41 and 42 show differences between estimates of mean TIMSS scale scores from the operational TIMSS samples and from the NAEP state samples using each of the three linkage methods. (See also Tables 1 and 2 from the ETS section and Tables 7 and 8 from the AIR section above.) The root mean square error (RMSE) provides an overall indicator of the accuracy of each linkage method in estimating state means. Confidence bounds for both the empirical TIMSS estimates and estimates using the NAEP linkages include estimates of sampling and measurement error. In addition, the estimates generated from the NAEP samples include error variance associated with error in estimating the linkage functions. Figures 5 and 6 show confidence bounds estimated for each of the empirical and linkage-based estimates of state means.

Table 41. Differences in Estimates of TIMSS Scale Score Means for Each Validation State - Mathematics

State	Actual	Mean Estimates using:			Projected Error (Predicted – Actual)		
	TIMSS	MOD	PRO	CAL	MOD	PRO	CAL
Nation	506.89	506.89	507.30	507.14	0.00	0.41	0.25
9-MA	560.58	540.00	538.10	540.34	-20.58	-22.48	-20.24
8-MN	544.73	532.52	531.63	533.22	-12.21	-13.09	-11.50
7-CO	517.79	525.80	525.35	526.20	8.00	7.56	8.40
6-CT	517.62	515.85	515.83	516.39	-1.78	-1.79	-1.24
5-NC	536.90	514.31	514.11	515.02	-22.59	-22.79	-21.87
4-IN	521.51	511.66	512.19	512.53	-9.85	-9.32	-8.98
3-FL	513.30	496.63	496.69	496.34	-16.68	-16.61	-16.97
2-CA	492.62	486.00	487.47	486.01	-6.62	-5.15	-6.61
1-AL	465.93	478.30	479.61	477.72	12.37	13.68	11.79
				Root Mean Square Error:	**13.83**	**14.27**	**13.51**

Note: MOD=Moderation; PRO=Projection; CAL=Calibration. The U.S. national samples for NAEP and TIMSS include students from both public and private schools. The “nation” results presented in the table were estimated using the students from public schools only, for comparison to the states which are restricted to public school students.

Table 42. Differences in Estimates of TIMSS Scale Score Means for Each Validation State - Science

State	Actual	Mean Estimates using:			Projected Error (Predicted – Actual)		
	TIMSS	MOD	PRO	CAL	MOD	PRO	CAL
Nation	522.19	522.19	522.43	522.29	0.00	0.24	0.10
9-MA	566.78	546.63	545.06	547.37	-20.15	-21.72	-19.41
8-MN	553.27	545.86	544.05	546.21	-7.41	-9.22	-7.07
7-CO	541.95	545.12	543.57	545.81	3.17	1.62	3.86
6-CT	531.60	531.34	531.32	531.53	-0.26	-0.28	-0.07
5-IN	532.80	527.35	526.77	527.14	-5.45	-6.03	-5.66
4-FL	529.89	516.71	517.66	516.98	-13.18	-12.23	-12.91
3-NC	531.53	515.16	516.37	514.53	-16.37	-15.17	-17.00
2-CA	498.52	498.12	499.96	498.25	-0.40	1.44	-0.27
1-AL	485.37	497.10	500.11	496.52	11.73	14.74	11.15
				Root Mean Square Error:	**10.95**	**11.52**	**10.82**

Note: MOD=Moderation; PRO=Projection; CAL=Calibration. The U.S. national samples for NAEP and TIMSS include students from both public and private schools. The “nation” results presented in the table were estimated using the students from public schools only, for comparison to the states which are restricted to public school students.

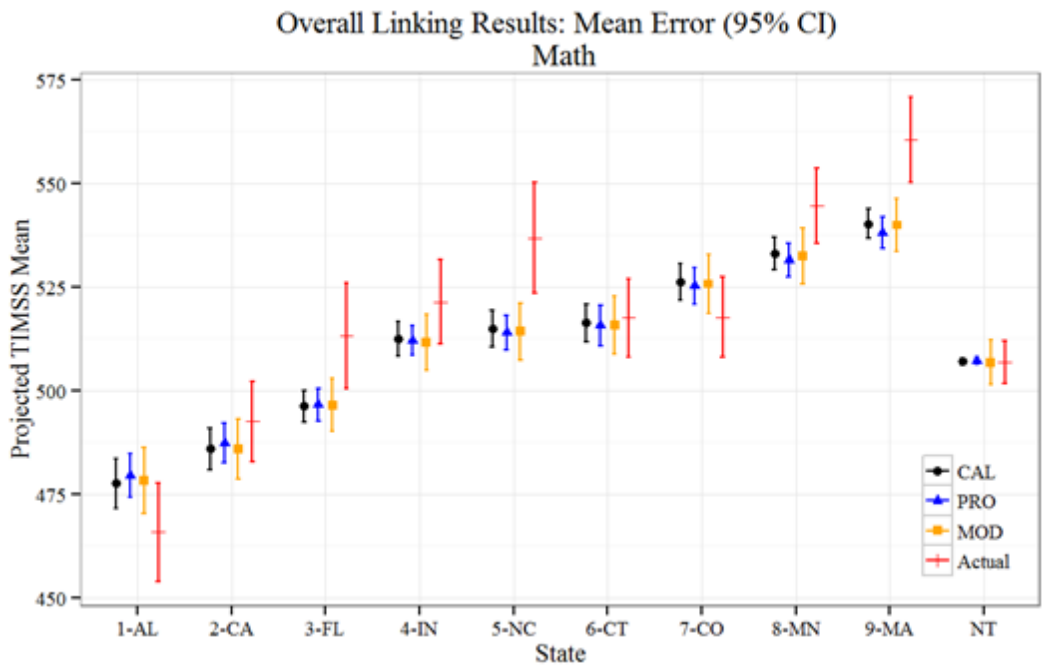


Figure 5. Confidence bounds for state mean estimates - Mathematics

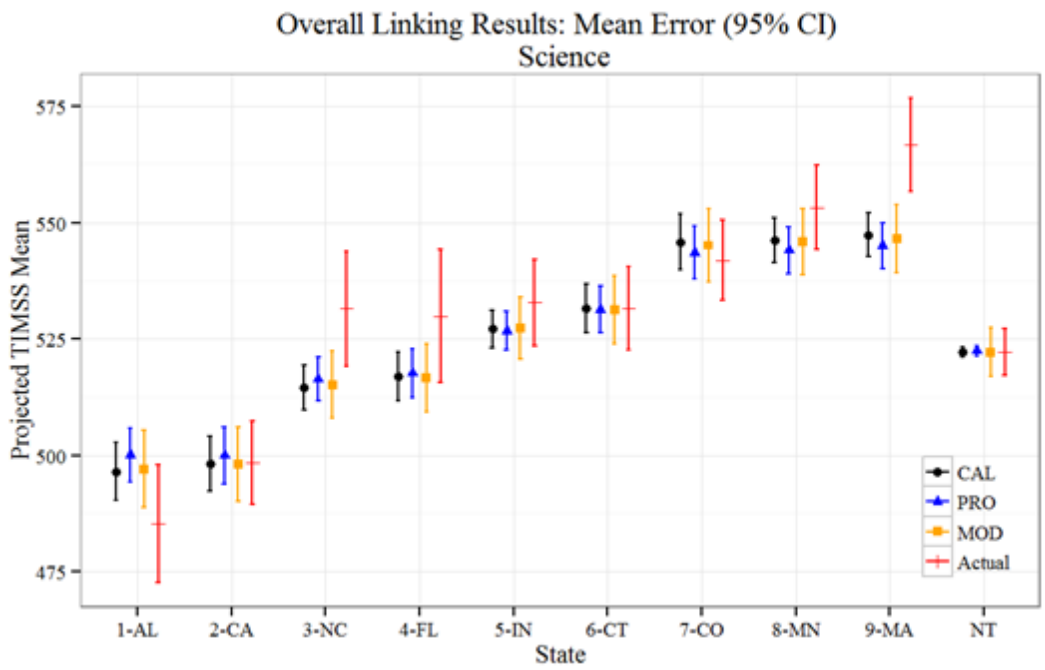


Figure 6. Confidence bounds for state mean estimates - Science

As shown in Figures 5 and 6 the confidence bounds for the empirical and linkage-based estimates of state means did not overlap for several states. Note that the confidence bounds for the empirical TIMSS means are larger than for the linkage-based projections because the TIMSS state samples are considerably smaller than the NAEP state samples used in generating the linkage-based projections.

Tables 43 and 44 show statistical significance of the difference between the empirical and linkage-based estimates for mathematics and science respectively. As shown, the differences were statistically significant for nearly half of the validation states in mathematics and for at least two validation states in science.

Tables 45 through 48 show differences in estimates of the percent above each of the TIMSS benchmark cutoffs along with statistical tests of these differences. As with the state means estimates, differences between empirical and linkage-based estimates were larger than would be expected based on estimates provided by AIR and ETS of the standard error of each estimate.

Table 43. Statistical Significance of Differences in Estimates of TIMSS Scale Score Means – Mathematics

			[A]MODERATION				PROJECTION				CALIBRATION			
	Actual TIMSS		Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	Mean	SE	Mean	SE	t	Sig.	Mean	SE	t	Sig.	Mean	SE	t	Sig.
Nation	506.89	2.63	506.89	2.75	0.00	1.000	507.30	0.43	0.16	0.877	507.14	0.45	0.09	0.925
9-MA	560.58	5.28	**540.00**	**3.29**	**-3.31**	**0.001**	**538.10**	**1.93**	**-4.00**	**0.000**	**540.34**	**1.78**	**-3.63**	**0.000**
8-MN	544.73	4.61	**532.52**	**3.45**	**-2.12**	**0.034**	**531.63**	**2.06**	**-2.59**	**0.010**	**533.22**	**1.98**	**-2.29**	**0.022**
7-CO	517.79	4.90	525.80	3.59	1.32	0.188	525.35	2.23	1.40	0.161	526.20	2.24	1.56	0.119
6-CT	517.62	4.84	515.85	3.55	-0.30	0.768	515.83	2.49	-0.33	0.742	516.39	2.35	-0.23	0.818
5-NC	536.90	6.85	**514.31**	**3.45**	**-2.94**	**0.003**	**514.11**	**2.14**	**-3.18**	**0.001**	**515.02**	**2.27**	**-3.03**	**0.002**
4-IN	521.51	5.13	511.66	3.42	-1.60	0.110	512.19	1.82	-1.71	0.087	512.53	2.13	-1.62	0.106
3-FL	513.30	6.45	**496.63**	**3.25**	**-2.31**	**0.021**	**496.69**	**1.97**	**-2.46**	**0.014**	**496.34**	**1.92**	**-2.52**	**0.012**
2-CA	492.62	4.88	486.00	3.73	-1.08	0.282	487.47	2.47	-0.94	0.347	486.01	2.60	-1.20	0.232
1-AL	465.93	5.93	478.30	4.05	1.70	0.090	**479.61**	**2.68**	**2.07**	**0.039**	477.72	3.04	1.74	0.082

[A:] Moderation results were based on moderation linking before the two-stage adjustment.
[B:] The standard error includes sampling and measurement errors only.
Note: Bold font indicates predicted means are statistically significant from the actual means. The U.S. national samples for NAEP and TIMSS include students from both public and private schools. The "nation" results presented in the table were estimated using the students from public schools only, for comparison to the states which are restricted to public school students.

Table 44. Statistical Significance of Differences in Estimates of TIMSS Scale Score Means – Science

			[A]MODERATION				PROJECTION				CALIBRATION			
	Actual TIMSS		Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	Mean	SE	Mean	SE	t	Sig.	Mean	SE	t	Sig.	Mean	SE	t	Sig.
Nation	522.19	2.53	522.19	2.71	0.00	1.000	522.43	0.55	0.09	0.926	522.29	0.55	0.04	0.970
9-MA	566.78	5.12	**546.63**	**3.73**	**-3.18**	**0.001**	**545.06**	**2.53**	**-3.80**	**0.000**	**547.37**	**2.37**	**-3.44**	**0.001**
8-MN	553.27	4.64	545.86	3.59	-1.26	0.207	544.05	2.54	-1.74	0.082	546.21	2.41	-1.35	0.177
7-CO	541.95	4.40	545.12	4.01	0.53	0.595	543.57	2.89	0.31	0.758	545.81	3.07	0.72	0.472
6-CT	531.60	4.57	531.34	3.73	-0.04	0.965	531.32	2.58	-0.05	0.957	531.53	2.67	-0.01	0.989
5-IN	532.80	4.75	527.35	3.39	-0.93	0.350	526.77	2.13	-1.16	0.247	527.14	2.07	-1.09	0.275
4-FL	529.89	7.30	516.71	3.75	-1.61	0.108	517.66	2.64	-1.57	0.115	516.98	2.71	-1.66	0.097
3-NC	531.53	6.28	**515.16**	**3.66**	**-2.25**	**0.024**	**516.37**	**2.39**	**-2.26**	**0.024**	**514.53**	**2.48**	**-2.52**	**0.012**
2-CA	498.52	4.56	498.12	4.10	-0.07	0.948	499.96	3.08	0.26	0.793	498.25	3.04	-0.05	0.961
1-AL	485.37	6.23	497.10	4.25	1.52	0.129	**500.11**	**2.98**	**2.07**	**0.038**	496.52	3.17	1.55	0.121

[A:] Moderation results were based on moderation linking before the two-stage adjustment.
[B:] The standard error includes sampling and measurement errors only.
Note: Bold font indicates predicted means are statistically significant from the actual means. The U.S. national samples for NAEP and TIMSS include students from both public and private schools. The "nation" results presented in the table were estimated using the students from public schools only, for comparison to the states which are restricted to public school students.

Table 45. Statistical Significance of Differences in Estimates of Percent Above Low TIMSS Achievement Level Cutoffs

Mathematics			[A]MODERATION			PROJECTION			CALIBRATION		
	Actual TIMSS		Projected		Error	Projected		[B]Error	Projected		[B]Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	97.72	0.34	96.62	0.67	-1.10	**96.20**	**0.57**	**-1.52**	**96.47**	**0.41**	**-1.25**
8-MN	97.17	0.67	95.32	0.86	-1.85	**95.17**	**0.60**	**-2.00**	**95.37**	**0.56**	**-1.80**
7-CO	93.48	1.07	94.67	1.03	1.18	94.57	0.56	1.09	94.65	0.59	1.17
6-CT	90.72	1.43	93.92	1.12	3.20	93.20	0.81	2.48	93.61	0.69	2.90
5-NC	95.34	1.31	93.42	1.24	-1.91	92.81	0.64	-2.53	93.13	0.72	-2.21
4-IN	95.07	0.96	94.49	1.18	-0.59	93.83	0.66	-1.24	94.17	0.68	-0.90
3-FL	93.76	1.31	90.32	1.41	-3.44	**89.37**	**0.75**	**-4.39**	**89.65**	**0.70**	**-4.12**
2-CA	87.45	1.72	85.36	1.84	-2.10	84.97	0.89	-2.48	84.85	0.84	-2.60
1-AL	78.61	2.32	**85.59**	**2.08**	**6.97**	**84.68**	**1.05**	**6.06**	**84.23**	**1.02**	**5.61**

Science			[A]MODERATION			PROJECTION			CALIBRATION		
	Actual TIMSS		Projected		Error	Projected		[B]Error	Projected		[B]Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	96.47	0.66	95.45	0.85	-1.02	95.19	0.65	-1.28	95.23	0.64	-1.24
8-MN	97.83	0.70	96.16	0.77	-1.67	**95.87**	**0.57**	**-1.96**	**95.87**	**0.55**	**-1.96**
7-CO	96.31	0.68	95.69	0.85	-0.62	95.64	0.57	-0.67	95.60	0.62	-0.72
6-CT	92.05	1.28	93.69	1.02	1.65	93.76	0.72	1.71	93.59	0.77	1.54
5-IN	95.11	0.86	94.28	1.00	-0.82	93.91	0.74	-1.20	94.13	0.75	-0.98
4-FL	93.48	1.49	91.28	1.39	-2.20	91.61	0.87	-1.87	91.29	0.98	-2.18
3-NC	94.37	1.38	92.13	1.39	-2.25	92.00	0.93	-2.38	91.81	0.81	-2.56
2-CA	87.53	1.64	86.13	1.78	-1.40	86.75	0.93	-0.77	86.39	1.01	-1.13
1-AL	83.39	1.91	87.76	2.04	4.36	**88.20**	**1.08**	**4.80**	87.37	1.05	3.98

[A:] Moderation results were based on moderation linking before the two-stage adjustment.
[B:] The standard error includes sampling and measurement errors only.
Note: P-A=Predicted minus Actual; Bold font indicates predicted estimates are statistically significant from the actual estimates.

Table 46. Statistical Significance of Differences in Estimates of Percent Above Intermediate TIMSS Achievement Level Cutoffs

Mathematics			[A]MODERATION			PROJECTION			CALIBRATION		
	Actual TIMSS		Projected		Error	Projected		[B]Error	Projected		[B]Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	88.07	1.39	**82.04**	**1.96**	**-6.04**	**81.35**	**1.19**	**-6.72**	**82.30**	**1.15**	**-5.78**
8-MN	82.75	1.86	79.44	2.17	-3.31	78.66	1.13	-4.09	79.55	1.10	-3.21
7-CO	70.58	2.53	75.87	2.43	5.29	75.64	1.14	5.06	75.95	1.10	5.37
6-CT	69.25	2.55	70.40	2.62	1.15	71.20	1.31	1.95	70.99	1.57	1.74
5-NC	77.90	2.51	**70.17**	**2.44**	**-7.73**	**70.01**	**1.24**	**-7.89**	**70.34**	**1.35**	**-7.57**
4-IN	74.13	2.34	71.16	2.67	-2.97	70.92	1.14	-3.21	71.38	1.23	-2.74
3-FL	67.60	3.31	62.20	2.50	-5.40	62.02	1.24	-5.58	62.01	1.05	-5.59
2-CA	59.04	2.76	56.11	2.68	-2.93	57.14	1.38	-1.90	56.49	1.28	-2.55
1-AL	45.76	3.20	53.60	3.15	7.85	**54.48**	**1.46**	**8.72**	**53.73**	**1.83**	**7.97**

Science			[A]MODERATION			PROJECTION			CALIBRATION		
	Actual TIMSS		Projected		[B]Error	Projected		Error	Projected		[B]Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	87.09	1.54	82.82	2.03	-4.27	**81.95**	**1.26**	**-5.14**	**82.68**	**1.01**	**-4.41**
8-MN	85.39	2.02	83.94	2.03	-1.46	82.85	1.13	-2.54	83.60	1.16	-1.79
7-CO	79.59	1.96	82.58	2.22	2.99	81.69	1.29	2.10	82.49	1.23	2.90
6-CT	74.23	2.00	77.66	2.40	3.43	77.09	1.21	2.86	77.02	1.34	2.79
5-IN	77.72	2.09	76.83	2.31	-0.89	76.37	1.17	-1.35	76.76	1.12	-0.96
4-FL	73.83	3.55	71.14	2.52	-2.70	71.50	1.56	-2.33	71.12	1.57	-2.71
3-NC	74.90	2.98	71.88	2.54	-3.02	71.73	1.47	-3.17	70.91	1.44	-3.99
2-CA	62.03	2.54	63.26	2.70	1.23	63.51	1.63	1.48	62.95	1.51	0.92
1-AL	56.20	3.73	64.61	3.08	8.41	**65.07**	**1.69**	**8.87**	63.77	1.71	7.57

[A:] Moderation results were based on moderation linking before the two-stage adjustment.
[B:] The standard error includes sampling and measurement errors only.
Note: P-A=Predicted minus Actual; Bold font indicates predicted estimates are statistically significant from the actual estimates.

Table 47. Statistical Significance of Differences in Estimates of Percent Above High TIMSS Achievement Level Cutoffs

Mathematics			[A]MODERATION			PROJECTION			CALIBRATION		
	Actual TIMSS		Projected		Error	Projected		[B]Error	Projected		[B]Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	57.35	3.22	**46.28**	**2.64**	**-11.07**	**45.53**	**1.31**	**-11.82**	**46.27**	**1.24**	**-11.08**
8-MN	48.90	2.84	42.55	2.73	-6.35	**42.40**	**1.29**	**-6.50**	43.30	1.26	-5.60
7-CO	35.14	2.69	38.70	2.69	3.56	38.96	1.45	3.82	39.25	1.42	4.11
6-CT	36.52	2.94	33.33	2.67	-3.20	33.73	1.43	-2.80	33.62	1.43	-2.91
5-NC	44.24	3.60	**32.40**	**2.48**	**-11.84**	**33.26**	**1.34**	**-10.97**	**33.08**	**1.30**	**-11.16**
4-IN	35.32	3.33	29.51	2.64	-5.81	30.82	1.21	-4.50	30.64	1.35	-4.68
3-FL	31.11	3.16	23.69	2.16	-7.41	24.95	1.03	-6.16	**24.44**	**1.12**	**-6.66**
2-CA	24.40	2.46	21.72	2.14	-2.68	22.93	1.18	-1.47	22.37	1.11	-2.03
1-AL	14.73	2.55	16.51	2.28	1.78	18.42	1.19	3.69	17.26	1.48	2.53

Science			[A]MODERATION			PROJECTION			CALIBRATION		
	Actual TIMSS		Projected		Error	Projected		[B]Error	Projected		[B]Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	61.46	2.79	**52.90**	**2.86**	**-8.57**	**50.76**	**1.75**	**-10.70**	**53.07**	**1.45**	**-8.40**
8-MN	53.67	2.62	52.23	3.04	-1.44	49.79	1.66	-3.88	52.11	1.50	-1.56
7-CO	47.86	2.58	51.34	3.35	3.48	49.47	1.63	1.60	51.42	1.94	3.56
6-CT	44.97	2.47	44.18	2.75	-0.79	43.43	1.70	-1.54	44.24	1.56	-0.73
5-IN	43.37	2.85	41.82	2.61	-1.55	40.83	1.36	-2.54	41.61	1.25	-1.76
4-FL	41.52	3.46	36.86	2.67	-4.65	36.82	1.51	-4.70	37.22	1.50	-4.30
3-NC	42.22	3.20	34.84	2.58	-7.37	**35.36**	**1.36**	**-6.86**	**34.90**	**1.37**	**-7.32**
2-CA	28.09	1.94	29.31	2.41	1.23	30.16	1.54	2.07	29.42	1.51	1.33
1-AL	23.77	2.76	27.14	2.51	3.37	28.74	1.38	4.98	27.44	1.54	3.67

[A:] Moderation results were based on moderation linking before the two-stage adjustment.
[B:] The standard error includes sampling and measurement errors only.
Note: P-A=Predicted minus Actual; Bold font indicates predicted estimates are statistically significant from the actual estimates.

Table 48. Statistical Significance of Differences in Estimates of Percent Above Advanced TIMSS Achievement Level Cutoffs

Mathematics			[A]MODERATION			PROJECTION			CALIBRATION		
	Actual TIMSS		Projected		Error	Projected		[B]Error	Projected		[B]Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	19.26	2.97	**11.33**	**1.69**	**-7.93**	**10.81**	**0.87**	**-8.45**	**11.44**	**0.77**	**-7.82**
8-MN	13.08	2.31	9.84	1.60	-3.25	9.46	0.74	-3.62	9.77	0.76	-3.32
7-CO	7.70	1.14	8.73	1.55	1.03	8.35	0.72	0.65	8.53	0.75	0.83
6-CT	10.17	1.34	6.93	1.31	-3.24	**6.81**	**0.78**	**-3.36**	7.26	0.66	-2.91
5-NC	13.75	2.63	**6.93**	**1.32**	**-6.82**	**6.67**	**0.68**	**-7.08**	**7.22**	**0.74**	**-6.53**
4-IN	6.98	1.18	4.38	1.12	-2.61	4.61	0.55	-2.37	4.57	0.53	-2.41
3-FL	7.92	1.59	**3.58**	**0.83**	**-4.34**	**3.96**	**0.43**	**-3.95**	**3.87**	**0.47**	**-4.05**
2-CA	4.82	0.91	4.40	1.06	-0.41	4.43	0.57	-0.39	4.55	0.67	-0.27
1-AL	2.10	0.77	1.91	0.67	-0.19	2.07	0.47	-0.03	2.08	0.46	-0.01

Science			[A]MODERATION			PROJECTION			CALIBRATION		
	Actual TIMSS		Projected		Error	Projected		[B]Error	Projected		[B]Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	24.46	2.55	**14.74**	**2.06**	**-9.72**	**15.18**	**0.96**	**-9.27**	**15.69**	**0.99**	**-8.76**
8-MN	16.13	1.87	12.49	1.97	-3.64	13.40	1.03	-2.73	13.38	1.04	-2.75
7-CO	14.46	1.62	13.68	2.09	-0.78	14.02	1.44	-0.44	14.77	1.25	0.31
6-CT	14.07	1.54	10.31	1.76	-3.76	11.08	1.03	-2.99	11.23	1.00	-2.84
5-IN	10.42	1.35	7.22	1.48	-3.20	8.66	0.79	-1.76	7.96	0.78	-2.46
4-FL	13.32	1.97	**7.34**	**1.34**	**-5.98**	**8.27**	**0.80**	**-5.05**	**7.82**	**0.78**	**-5.50**
3-NC	12.42	2.18	**6.51**	**1.34**	**-5.92**	**7.58**	**0.74**	**-4.85**	**6.96**	**0.77**	**-5.46**
2-CA	6.03	0.73	5.76	1.32	-0.27	6.58	0.74	0.55	6.27	0.79	0.24
1-AL	4.81	1.01	3.64	1.14	-1.17	5.06	0.79	0.25	4.19	0.64	-0.62

[A]: Moderation results were based on moderation linking before the two-stage adjustment.
[B]: The standard error includes sampling and measurement errors only.
Note: P-A=Predicted minus Actual Bold font indicates predicted estimates are statistically significant from the actual estimates.

Tables 49 and 50 show differences for each gender between estimates of mean TIMSS scale scores from the operational TIMSS and predicted TIMSS means from each of the three linkage methods. At the national level, the errors for each gender were small and not statistically significant, although the PRO method yielded slightly larger errors (greater than half a scale score) in the estimates for males compared to the other two methods. The pattern of statistically significant differences at the state-level was similar for males and females, both following the pattern of overall errors in state level mean estimates. Figures 7 and 8 display the confidence bounds for the empirical linkage-based estimates of state means for both males and females.

Table 49. Statistical Significance of Differences in Estimates of TIMSS Scale Score Means for Males and Females – Mathematics

Math-Males			[A]MODERATION				PROJECTION				CALIBRATION			
	Actual TIMSS		Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	Mean	SE	Mean	SE	t	Sig.	Mean	SE	t	Sig.	Mean	SE	t	Sig.
Nation	507.97	2.82	507.74	2.78	-0.06	0.954	508.10	0.59	0.04	0.964	508.01	0.61	0.01	0.990
9-MA	563.26	5.50	**540.77**	**3.76**	**-3.38**	**0.001**	**539.02**	**2.78**	**-3.93**	**0.000**	**541.16**	**2.66**	**-3.62**	**0.000**
8-MN	544.90	5.12	**531.70**	**3.54**	**-2.12**	**0.034**	**530.73**	**2.60**	**-2.47**	**0.014**	**532.55**	**2.26**	**-2.21**	**0.027**
7-CO	519.60	4.95	524.98	3.76	0.87	0.387	524.80	2.66	0.93	0.355	525.13	2.50	1.00	0.319
6-CT	515.62	5.45	518.21	4.05	0.38	0.702	517.75	3.03	0.34	0.732	518.49	2.84	0.47	0.640
5-NC	538.54	8.38	**512.57**	**4.05**	**-2.79**	**0.005**	**512.82**	**2.98**	**-2.89**	**0.004**	**513.44**	**3.07**	**-2.81**	**0.005**
4-IN	525.59	5.88	512.00	3.77	-1.95	0.052	512.95	2.69	-1.95	0.051	513.54	2.70	-1.86	0.063
3-FL	517.07	7.33	**497.51**	**3.41**	**-2.42**	**0.015**	**497.16**	**2.32**	**-2.59**	**0.010**	**496.70**	**2.29**	**-2.65**	**0.008**
2-CA	494.32	5.04	486.24	4.24	-1.23	0.220	487.84	3.13	-1.09	0.275	486.30	3.33	-1.33	0.184
1-AL	465.10	6.33	478.40	4.33	1.73	0.083	**479.92**	**3.24**	**2.08**	**0.037**	477.96	3.73	1.75	0.080

Math-Females			[A]MODERATION				PROJECTION				CALIBRATION			
	Actual TIMSS		Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	Mean	SE	Mean	SE	t	Sig.	Mean	SE	t	Sig.	Mean	SE	t	Sig.
Nation	505.82	2.89	506.01	2.76	0.05	0.963	506.48	0.51	0.22	0.823	506.24	0.50	0.14	0.886
9-MA	557.94	5.96	**539.20**	**3.42**	**-2.73**	**0.006**	**537.15**	**2.31**	**-3.26**	**0.001**	**539.50**	**1.94**	**-2.94**	**0.003**
8-MN	544.56	4.90	533.37	3.96	-1.78	0.076	**532.57**	**2.44**	**-2.19**	**0.029**	533.92	2.66	-1.91	0.057
7-CO	516.07	5.38	526.63	4.06	1.57	0.117	525.91	2.80	1.62	0.105	527.30	2.97	1.83	0.068
6-CT	519.68	5.21	513.50	3.74	-0.96	0.335	513.93	2.92	-0.96	0.335	514.29	2.62	-0.92	0.356
5-NC	535.36	6.21	**516.11**	**3.52**	**-2.69**	**0.007**	**515.45**	**2.31**	**-3.00**	**0.003**	**516.66**	**2.40**	**-2.81**	**0.005**
4-IN	517.76	5.10	511.32	3.67	-1.02	0.306	511.44	2.19	-1.14	0.255	511.53	2.43	-1.10	0.271
3-FL	509.31	6.65	495.72	3.58	-1.80	0.072	496.21	2.47	-1.85	0.065	495.96	2.38	-1.89	0.059
2-CA	490.88	5.55	485.74	4.02	-0.75	0.454	487.08	3.03	-0.60	0.548	485.70	2.93	-0.82	0.410
1-AL	466.72	6.41	478.18	4.29	1.49	0.137	479.29	3.28	1.75	0.081	477.47	3.18	1.50	0.133

[A]: Moderation results were based on moderation linking before the two-stage adjustment.
[B]: The standard error includes sampling and measurement errors only.
Note: Bold font indicates predicted means are statistically significant from the actual means.

Table 50. Statistical Significance of Differences in Estimates of TIMSS Scale Score Means for Males and Females – Science

Science-Males			[A]MODERATION				PROJECTION				CALIBRATION			
	Actual TIMSS		Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	Mean	SE	Mean	SE	t	Sig.	Mean	SE	t	Sig.	Mean	SE	t	Sig.
Nation	527.39	2.81	527.25	2.74	-0.03	0.972	527.05	0.72	-0.12	0.906	527.40	0.64	0.00	0.998
9-MA	570.09	5.06	**552.51**	**4.20**	**-2.67**	**0.008**	**550.58**	**3.21**	**-3.26**	**0.001**	**553.41**	**3.18**	**-2.79**	**0.005**
8-MN	559.35	5.29	551.60	3.84	-1.19	0.235	549.18	3.00	-1.67	0.094	552.05	2.77	-1.22	0.222
7-CO	547.65	5.13	548.90	4.26	0.19	0.851	547.23	3.24	-0.07	0.946	550.17	3.23	0.42	0.677
6-CT	532.91	5.86	534.62	4.29	0.24	0.814	534.70	3.35	0.26	0.791	535.44	3.47	0.37	0.710
5-IN	540.52	5.40	536.09	3.84	-0.67	0.504	534.99	3.28	-0.87	0.382	535.54	2.92	-0.81	0.417
4-FL	536.95	7.57	**519.37**	**4.31**	**-2.02**	**0.043**	**520.30**	**3.68**	**-1.98**	**0.048**	**519.40**	**3.69**	**-2.08**	**0.037**
3-NC	537.49	7.72	**517.86**	**4.25**	**-2.23**	**0.026**	**518.68**	**3.49**	**-2.22**	**0.026**	**517.33**	**3.16**	**-2.42**	**0.016**
2-CA	504.30	5.03	503.18	4.80	-0.16	0.872	504.52	3.95	0.04	0.972	503.02	3.88	-0.20	0.841
1-AL	488.85	6.94	499.27	4.68	1.25	0.213	501.41	3.81	1.59	0.113	498.95	3.81	1.28	0.202

Science-Females			[A]MODERATION				PROJECTION				CALIBRATION			
	Actual TIMSS		Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	Mean	SE	Mean	SE	t	Sig.	Mean	SE	t	Sig.	Mean	SE	t	Sig.
Nation	517.09	2.74	516.95	2.75	-0.03	0.972	517.66	0.73	0.20	0.840	517.00	0.73	-0.03	0.977
9-MA	563.51	5.78	**540.57**	**4.21**	**-3.21**	**0.001**	**539.36**	**3.23**	**-3.64**	**0.000**	**541.15**	**3.06**	**-3.42**	**0.001**
8-MN	547.61	4.92	539.91	4.08	-1.20	0.228	538.73	3.19	-1.51	0.130	540.14	3.09	-1.29	0.199
7-CO	536.51	4.70	541.22	4.92	0.69	0.489	539.79	4.10	0.53	0.599	541.31	4.18	0.76	0.445
6-CT	530.25	4.48	528.06	4.04	-0.36	0.716	527.94	3.19	-0.42	0.674	527.60	3.07	-0.49	0.626
5-IN	525.72	4.88	518.62	3.77	-1.15	0.249	518.56	2.85	-1.27	0.205	518.75	2.69	-1.25	0.211
4-FL	522.42	8.48	513.96	4.32	-0.89	0.374	514.93	3.44	-0.82	0.413	514.46	3.30	-0.87	0.382
3-NC	525.94	5.72	512.39	4.06	-1.93	0.053	514.00	2.95	-1.86	0.063	**511.66**	**3.00**	**-2.21**	**0.027**
2-CA	492.57	4.96	492.77	4.37	0.03	0.976	495.13	3.55	0.42	0.674	493.21	3.48	0.11	0.916
1-AL	482.03	6.50	494.86	4.76	1.59	0.111	**498.78**	**3.73**	**2.23**	**0.026**	494.02	3.76	1.60	0.110

[A]: Moderation results were based on moderation linking before the two-stage adjustment.
[B]: The standard error includes sampling and measurement errors only.
Note: Bold font indicates predicted means are statistically significant from the actual means.

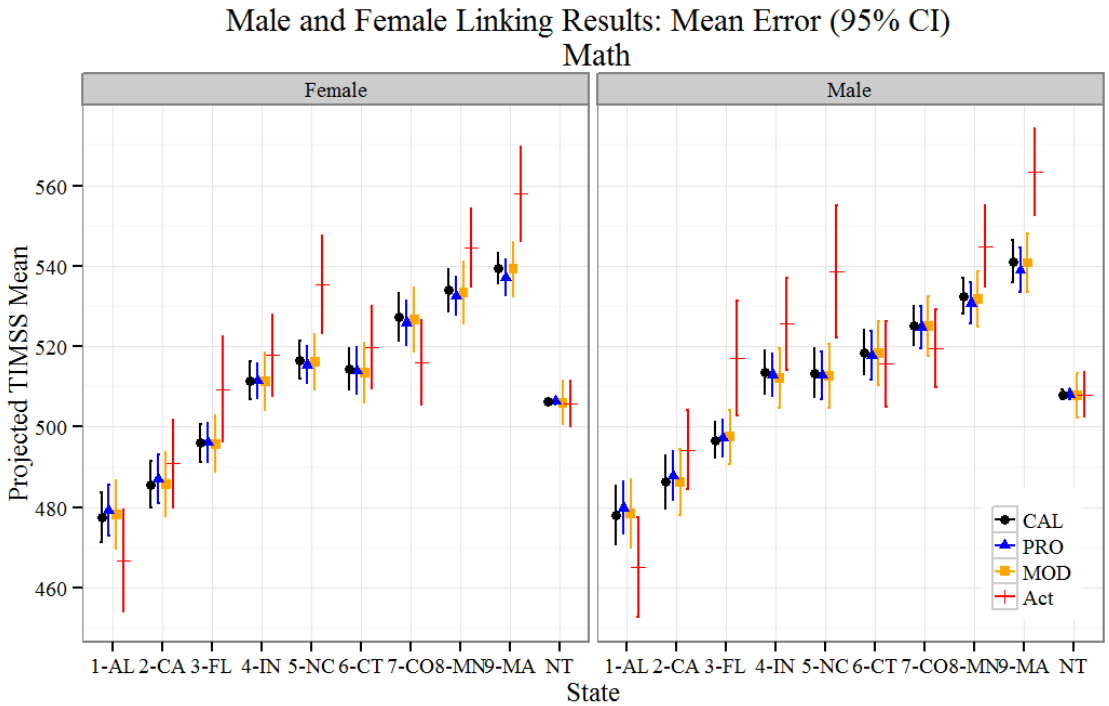


Figure 7. Confidence Bounds for State Mean Estimates for Males and Females – Mathematics.

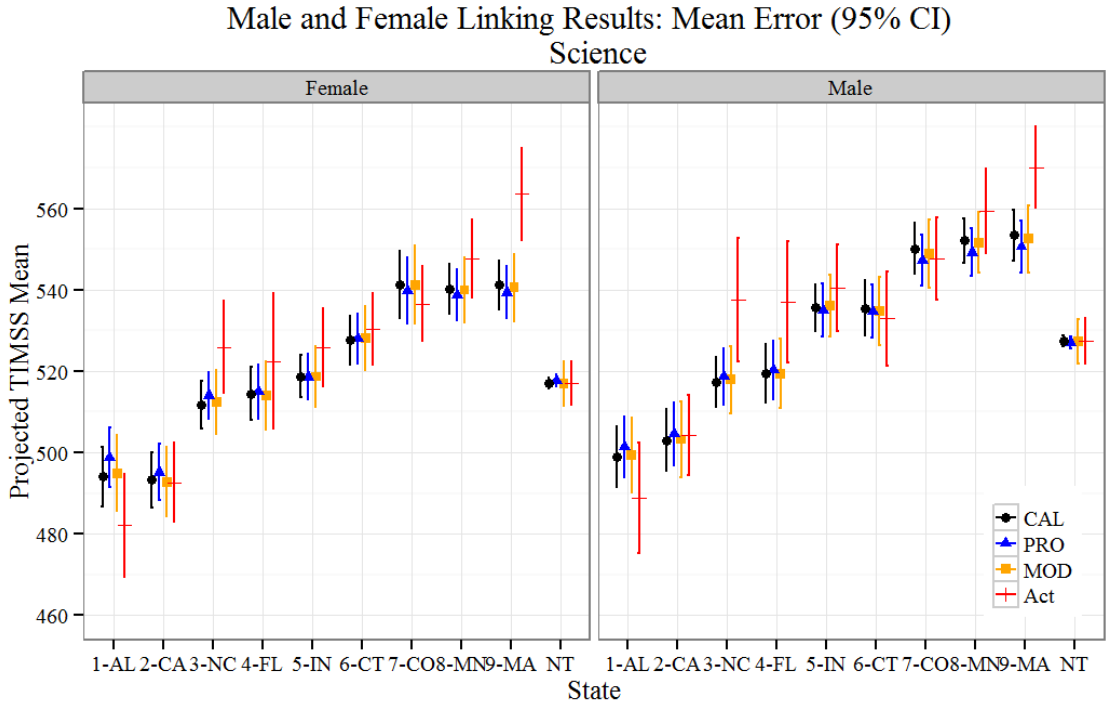


Figure 8. Confidence Bounds for State Mean Estimates for Males and Females – Science.

This NCES document was prepared to supplement the "Linking Methodologies" section presented in the Global Context report (NCES 2013-460).

Tables 51 through 58 show differences between estimates of mean TIMSS scale scores from the operational TIMSS and predicted TIMSS means from each of the three linkage methods for each of these racial/ethnic groups. At the national level, estimation errors for some groups, while not statistically significant, were quite a bit larger in comparison to estimation errors by gender (several scale score points compared to less than one). Again, the pattern of differences for each racial/ethnic group at the state level was similar to the pattern of errors in the overall state mean estimates. Note that the projection method, which accounted for some demographic information, yielded far smaller differences by race/ethnicity compared to the other two methods.

Table 51. Statistical Significance of Differences in Estimates of TIMSS Scale Score Means for Whites – Mathematics

Math-White			[A]MODERATION				PROJECTION				CALIBRATION			
	Actual TIMSS		Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	Mean	SE	Mean	SE	t	Sig.	Mean	SE	t	Sig.	Mean	SE	t	Sig.
Nation	528.29	2.94	530.56	2.81	0.56	0.578	529.61	0.50	0.44	0.659	531.10	0.48	0.94	0.346
9-MA	572.04	5.54	**554.57**	**3.42**	**-2.68**	**0.007**	**552.02**	**2.12**	**-3.38**	**0.001**	**555.22**	**1.74**	**-2.90**	**0.004**
8-MN	557.59	4.60	550.62	3.44	-1.21	0.225	548.26	1.88	-1.88	0.060	550.90	1.75	-1.36	0.174
7-CO	544.10	5.22	549.58	3.82	0.85	0.397	546.74	2.32	0.46	0.643	549.90	2.31	1.02	0.309
6-CT	543.23	5.52	540.80	3.58	-0.37	0.712	539.21	2.45	-0.66	0.507	541.41	2.18	-0.31	0.760
5-NC	563.42	7.31	**535.89**	**3.59**	**-3.38**	**0.001**	**534.71**	**2.48**	**-3.72**	**0.000**	**537.10**	**2.34**	**-3.43**	**0.001**
4-IN	530.44	5.66	524.79	3.52	-0.85	0.397	524.32	2.01	-1.02	0.308	525.55	2.18	-0.81	0.420
3-FL	530.93	6.10	521.21	3.83	-1.35	0.177	519.88	3.02	-1.62	0.104	521.19	3.07	-1.43	0.154
2-CA	525.06	6.42	523.26	5.48	-0.21	0.831	522.59	4.68	-0.31	0.756	524.03	4.71	-0.13	0.897
1-AL	489.18	6.72	502.48	4.27	1.67	0.095	502.85	3.25	1.83	0.067	502.86	3.56	1.80	0.072

[A]: Moderation results were based on moderation linking before the two-stage adjustment.
[B]: The standard error includes sampling and measurement errors only.
Note: Bold font indicates predicted means are statistically significant from the actual means.

Table 52. Statistical Significance of Differences in Estimates of TIMSS Scale Score Means for African Americans – Mathematics

Math-African-American			[A]MODERATION				PROJECTION				CALIBRATION			
	Actual TIMSS		Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	Mean	SE	Mean	SE	t	Sig.	Mean	SE	t	Sig.	Mean	SE	t	Sig.
Nation	468.21	4.12	463.42	3.04	-0.94	0.349	465.88	0.91	-0.55	0.581	462.96	0.96	-1.24	0.214
9-MA	516.44	8.57	499.42	7.64	-1.48	0.138	500.75	7.46	-1.38	0.167	500.37	6.79	-1.47	0.142
8-MN	497.03	12.27	470.22	7.16	-1.89	0.059	473.80	7.97	-1.59	0.112	471.74	6.98	-1.79	0.073
7-CO	486.53	21.70	482.24	7.55	-0.19	0.852	483.23	7.63	-0.14	0.886	482.49	7.16	-0.18	0.860
6-CT	452.54	10.36	473.78	5.25	1.83	0.067	476.37	6.65	1.94	0.053	473.41	5.10	1.81	0.071
5-NC	494.56	8.52	476.28	4.57	-1.89	0.059	477.34	3.52	-1.87	0.062	**475.72**	**3.61**	**-2.04**	**0.042**
4-IN	467.13	9.54	467.28	6.44	0.01	0.989	469.82	5.40	0.25	0.806	467.96	5.73	0.07	0.940
3-FL	484.02	8.18	**456.29**	**4.93**	**-2.90**	**0.004**	**458.08**	**4.48**	**-2.78**	**0.005**	**455.12**	**3.69**	**-3.22**	**0.001**
2-CA	467.72	12.48	439.30	8.39	-1.89	0.059	444.32	7.47	-1.61	0.107	440.78	7.63	-1.84	0.065
1-AL	427.94	4.86	439.15	4.76	1.65	0.099	**441.60**	**3.78**	**2.22**	**0.027**	437.07	3.83	1.47	0.140

[A]: Moderation results were based on moderation linking before the two-stage adjustment.
[B]: The standard error includes sampling and measurement errors only.
Note: Bold font indicates predicted means are statistically significant from the actual means.

Table 53. Statistical Significance of Differences in Estimates of TIMSS Scale Score Means for Hispanics – Mathematics

Math-Hispanic			[A]MODERATION				PROJECTION				CALIBRATION			
	Actual TIMSS		Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	Mean	SE	Mean	SE	t	Sig.	Mean	SE	t	Sig.	Mean	SE	t	Sig.
Nation	482.26	3.38	480.04	2.90	-0.50	0.618	482.04	0.94	-0.06	0.949	479.82	0.84	-0.70	0.484
9-MA	507.11	7.11	490.92	4.24	-1.95	0.051	492.07	4.02	-1.84	0.066	**490.44**	**3.91**	**-2.05**	**0.040**
8-MN	495.56	5.73	485.27	5.28	-1.32	0.187	487.47	5.35	-1.03	0.302	486.09	4.89	-1.26	0.209
7-CO	480.43	5.12	486.79	4.07	0.97	0.331	490.38	3.13	1.66	0.097	487.55	3.02	1.20	0.231
6-CT	467.12	6.13	468.22	4.43	0.15	0.884	471.43	3.77	0.60	0.549	468.69	3.72	0.22	0.826
5-NC	509.54	9.29	**489.59**	**4.07**	**-1.97**	**0.049**	491.21	3.77	-1.83	0.067	490.68	3.77	-1.88	0.060
4-IN	500.59	7.20	484.97	4.53	-1.84	0.066	487.88	4.66	-1.48	0.138	485.86	3.96	-1.79	0.073
3-FL	505.40	9.46	486.24	3.20	-1.92	0.055	486.93	2.06	-1.91	0.056	**485.73**	**1.82**	**-2.04**	**0.041**
2-CA	470.00	5.58	461.77	3.58	-1.24	0.215	464.40	2.41	-0.92	0.357	461.11	2.10	-1.49	0.136
1-AL	454.38	9.54	446.21	6.42	-0.71	0.477	449.76	6.34	-0.40	0.687	444.50	6.41	-0.86	0.390

[A]: Moderation results were based on moderation linking before the two-stage adjustment.
[B]: The standard error includes sampling and measurement errors only.
Note: Bold font indicates predicted means are statistically significant from the actual means.

Table 54. Statistical Significance of Differences in Estimates of TIMSS Scale Score Means for Asians – Mathematics

Math-Asian			[A]MODERATION				PROJECTION				CALIBRATION			
	Actual TIMSS		Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	Mean	SE	Mean	SE	t	Sig.	Mean	SE	t	Sig.	Mean	SE	t	Sig.
Nation	560.44	7.25	557.82	3.65	-0.32	0.748	554.56	1.98	-0.78	0.435	558.78	2.08	-0.22	0.826
9-MA	599.08	7.95	578.85	7.28	-1.88	0.060	**571.14**	**8.81**	**-2.35**	**0.019**	**578.07**	**6.78**	**-2.01**	**0.044**
8-MN	536.29	17.32	508.53	9.24	-1.41	0.157	509.14	8.37	-1.41	0.158	509.94	9.12	-1.35	0.178
7-CO	545.13	12.03	570.73	9.14	1.69	0.090	567.26	8.88	1.48	0.139	570.47	8.70	1.71	0.088
6-CT	576.76	12.20	561.62	8.64	-1.01	0.311	556.05	8.47	-1.39	0.163	559.77	7.56	-1.18	0.237
5-NC	604.77	16.69	570.22	10.76	-1.74	0.082	**563.83**	**12.16**	**-1.98**	**0.047**	570.44	11.22	-1.71	0.088
4-IN	521.22	26.47	559.19	16.34	1.22	0.222	552.94	14.85	1.05	0.296	560.69	12.45	1.35	0.177
3-FL	614.80	15.09	**569.28**	**9.36**	**-2.56**	**0.010**	**565.69**	**8.95**	**-2.80**	**0.005**	**570.99**	**9.23**	**-2.48**	**0.013**
2-CA	555.33	9.48	550.81	6.53	-0.39	0.694	548.80	5.52	-0.60	0.552	551.78	5.31	-0.33	0.744
1-AL	509.35	32.89	533.66	13.25	0.69	0.493	531.80	14.04	0.63	0.530	533.23	13.41	0.67	0.501

[A]: Moderation results were based on moderation linking before the two-stage adjustment.
[B]: The standard error includes sampling and measurement errors only.
Note: Bold font indicates predicted means are statistically significant from the actual means.

Table 55. Statistical Significance of Differences in Estimates of TIMSS Scale Score Means for Whites – Science

Science-White			[A]MODERATION				PROJECTION				CALIBRATION			
	Actual TIMSS		Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	Mean	SE	Mean	SE	t	Sig.	Mean	SE	t	Sig.	Mean	SE	t	Sig.
Nation	551.60	2.82	553.84	2.79	0.57	0.572	551.45	0.54	-0.05	0.957	554.52	0.57	1.01	0.310
9-MA	586.62	5.10	**569.91**	**3.64**	**-2.67**	**0.008**	**566.49**	**2.44**	**-3.56**	**0.000**	**571.07**	**2.04**	**-2.83**	**0.005**
8-MN	569.62	4.25	565.76	3.65	-0.69	0.490	562.38	2.68	-1.44	0.149	566.15	2.17	-0.73	0.467
7-CO	572.00	4.29	572.42	4.19	0.07	0.944	568.25	3.16	-0.70	0.482	573.80	3.32	0.33	0.740
6-CT	561.55	5.06	560.14	3.68	-0.23	0.821	557.71	2.41	-0.68	0.494	560.72	2.53	-0.15	0.883
5-IN	546.49	5.28	547.17	3.56	0.11	0.916	545.07	2.72	-0.24	0.811	547.27	2.28	0.13	0.893
4-FL	560.39	6.10	549.87	4.03	-1.44	0.150	547.95	3.02	-1.83	0.067	550.11	3.09	-1.50	0.133
3-NC	564.72	6.36	**544.57**	**3.80**	**-2.72**	**0.007**	**543.24**	**2.66**	**-3.11**	**0.002**	**544.51**	**3.04**	**-2.87**	**0.004**
2-CA	545.99	6.63	548.64	5.82	0.30	0.764	546.64	4.85	0.08	0.937	549.74	5.43	0.44	0.662
1-AL	518.81	5.52	527.79	4.30	1.28	0.199	528.44	3.02	1.53	0.125	527.62	3.02	1.40	0.161

[A]: Moderation results were based on moderation linking before the two-stage adjustment.
[B]: The standard error includes sampling and measurement errors only.
Note: Bold font indicates predicted means are statistically significant from the actual means.

Table 56. Statistical Significance of Differences in Estimates of TIMSS Scale Score Means for African-Americans – Science

Science-African-American			[A]MODERATION				PROJECTION				CALIBRATION			
	Actual TIMSS		Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	Mean	SE	Mean	SE	t	Sig.	Mean	SE	t	Sig.	Mean	SE	t	Sig.
Nation	473.44	4.02	468.55	3.06	-0.97	0.333	473.48	1.20	0.01	0.991	467.86	1.18	-1.33	0.184
9-MA	514.05	9.92	490.95	9.44	-1.69	0.092	493.20	8.92	-1.56	0.118	492.39	8.96	-1.62	0.105
8-MN	488.50	13.19	465.79	7.10	-1.52	0.129	469.93	6.94	-1.25	0.213	464.14	7.24	-1.62	0.105
7-CO	507.39	18.80	506.68	11.60	-0.03	0.975	509.29	11.48	0.09	0.931	505.32	9.92	-0.10	0.922
6-CT	458.53	10.92	467.56	6.70	0.70	0.481	473.29	6.46	1.16	0.245	465.87	7.01	0.57	0.572
5-IN	460.48	9.80	466.34	7.71	0.47	0.638	470.17	7.11	0.80	0.423	465.50	7.92	0.40	0.690
4-FL	484.93	9.93	465.50	5.89	-1.68	0.092	470.67	4.99	-1.28	0.200	466.63	5.23	-1.63	0.103
3-NC	481.34	6.48	**463.32**	**5.25**	**-2.16**	**0.031**	468.86	4.61	-1.57	0.117	**461.51**	**4.18**	**-2.57**	**0.010**
2-CA	459.52	12.56	455.76	9.41	-0.24	0.811	461.98	9.60	0.16	0.876	454.36	9.67	-0.33	0.745
1-AL	435.17	5.24	446.29	4.67	1.59	0.113	**453.57**	**3.65**	**2.88**	**0.004**	445.48	3.98	1.57	0.117

[A]: Moderation results were based on moderation linking before the two-stage adjustment.
[B]: The standard error includes sampling and measurement errors only.
Note: Bold font indicates predicted means are statistically significant from the actual means.

Table 57. Statistical Significance of Differences in Estimates of TIMSS Scale Score Means for Hispanics – Science

Science-Hispanic			[A]MODERATION				PROJECTION				CALIBRATION			
	Actual TIMSS		Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	Mean	SE	Mean	SE	t	Sig.	Mean	SE	t	Sig.	Mean	SE	t	Sig.
Nation	491.31	3.39	490.73	2.92	-0.13	0.897	493.51	1.20	0.61	0.540	490.02	1.04	-0.36	0.717
9-MA	493.69	9.40	486.43	6.07	-0.65	0.517	490.00	5.59	-0.34	0.736	485.25	5.70	-0.77	0.443
8-MN	511.96	7.17	497.39	6.70	-1.49	0.137	499.67	6.99	-1.23	0.219	498.78	6.65	-1.35	0.178
7-CO	499.35	5.26	505.84	4.57	0.93	0.352	507.99	3.94	1.31	0.189	505.66	4.49	0.91	0.362
6-CT	474.37	5.28	481.76	5.68	0.95	0.340	485.65	4.86	1.57	0.116	481.85	4.42	1.09	0.277
5-IN	498.58	6.16	489.88	6.55	-0.97	0.333	492.54	6.18	-0.69	0.488	489.07	5.24	-1.18	0.240
4-FL	523.18	10.28	505.58	4.48	-1.57	0.116	507.37	4.09	-1.43	0.153	505.13	3.61	-1.66	0.097
3-NC	502.11	8.68	491.94	6.04	-0.96	0.336	494.93	5.36	-0.70	0.481	491.38	5.77	-1.03	0.303
2-CA	474.94	5.35	471.94	4.01	-0.45	0.653	475.72	3.13	0.13	0.900	471.43	2.81	-0.58	0.561
1-AL	469.75	9.85	469.87	7.73	0.01	0.992	474.49	7.18	0.39	0.697	467.31	8.01	-0.19	0.848

[A]: Moderation results were based on moderation linking before the two-stage adjustment.
[B]: The standard error includes sampling and measurement errors only.
Note: Bold font indicates predicted means are statistically significant from the actual means.

Table 58. Statistical Significance of Differences in Estimates of TIMSS Scale Score Means for Asians – Science

Science-Asian			[A]MODERATION				PROJECTION				CALIBRATION			
	Actual TIMSS		Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	Mean	SE	Mean	SE	t	Sig.	Mean	SE	t	Sig.	Mean	SE	t	Sig.
Nation	547.67	7.13	548.70	4.07	0.13	0.900	546.41	2.81	-0.16	0.870	550.33	2.93	0.35	0.729
9-MA	576.06	8.80	567.02	10.85	-0.65	0.518	564.05	11.12	-0.85	0.397	569.37	9.32	-0.52	0.602
8-MN	511.36	13.93	516.10	8.81	0.29	0.773	516.66	8.79	0.32	0.747	515.82	8.17	0.28	0.782
7-CO	548.85	14.75	543.11	12.49	-0.30	0.766	543.24	13.59	-0.28	0.780	542.08	12.08	-0.36	0.722
6-CT	565.24	13.82	559.56	9.13	-0.34	0.732	556.24	8.95	-0.55	0.585	559.33	10.17	-0.34	0.731
5-IN	492.42	26.87	550.88	16.74	1.85	0.065	546.74	18.00	1.68	0.093	551.63	19.65	1.78	0.075
4-FL	600.13	14.01	**562.63**	**8.70**	**-2.27**	**0.023**	**560.30**	**10.35**	**-2.29**	**0.022**	**565.27**	**7.38**	**-2.20**	**0.028**
3-NC	576.74	17.85	544.86	15.19	-1.36	0.174	543.05	14.43	-1.47	0.142	545.76	16.35	-1.28	0.201
2-CA	542.48	9.11	542.23	8.00	-0.02	0.984	540.54	6.94	-0.17	0.866	544.22	7.38	0.15	0.882
1-AL	493.14	35.41	502.70	15.42	0.25	0.805	506.74	13.29	0.36	0.719	503.36	16.20	0.26	0.793

[A]: Moderation results were based on moderation linking before the two-stage adjustment.
[B]: The standard error includes sampling and measurement errors only.
Note: Bold font indicates predicted means are statistically significant from the actual means.

Results from the comparisons of empirical and linkage-based estimates led to the following general conclusions:

Finding 1: The three different linkage methods yield similar linkage functions.

In all cases, differences in the estimates produced by the three different linkage methods are quite small in comparison to differences between each of the linkage-based estimates and the empirical TIMSS results.

Finding 2: Confidence bounds for each of the linkage-based estimates omit significant sources of error.

Estimates of sampling and measurement error for both the NAEP and TIMSS samples are well established. Linking function error for the statistical moderation approach is based on well-established estimates of variation in the national NAEP and TIMSS means and standard deviations. Observed differences between the empirical and linkage-based estimates are larger than predicted by these sources of variation, so other differences between the two assessments must be contributing significant amounts of variation in the estimates.

Finding 3: The three different linkage methods yield similar linkage functions at national subgroups.

The difference between predicted and actual TIMSS results was not statistically significant for any national gender or racial/ethnic group across all linking methods. The pattern of statistically significant differences at the state-level was similar for males and females, both following the pattern of overall errors in state level mean estimates. Again, the pattern of differences for each racial/ethnic group at the state level was similar to the pattern of errors in the overall state mean estimates.

Primary Conclusions – Stage 2

HumRRO investigated the impact of two key differences between the NAEP and TIMSS assessments: (1) differences in exclusion and accommodation policies, and (2) differences in the distribution of test item difficulty and item formats. Other differences, such as the difference in testing window, could not be investigated within the scope of the current study.

Differences in Accommodation and Exclusion Rates

Tables 59 and 60 show the percentage of students in each of the validation states excluded from the NAEP and TIMSS assessments and the percentage receiving one or more testing accommodations in the NAEP assessment for mathematics and science respectively[7]. The NAEP

[7] Note that TIMSS combines the mathematics and science assessments, so exclusion rates are the same for these two subjects.

program has worked assiduously in recent years to maximize inclusion rates by offering a menu of accommodations and ensuring states and schools correctly include students who can be accommodated. Over time, in general NAEP accommodation rates have grown while exclusion rates declined. However, NAEP exclusion and accommodation rates varied considerably across the nine validation states. TIMSS allows few, if any, accommodations and data on TIMSS accommodation rates were not available. As shown, TIMSS exclusion rates are considerably higher than NAEP exclusion rates. The difference between the percent excluded in the NAEP and TIMSS assessments also varies considerably from state to state.

Table 59. NAEP and TIMSS Exclusion and Accommodation Rates—Mathematics

2011 NAEP/TIMSS Math: Exclusion & Accommodation Percentages					
State	NAEP			TIMSS	Diff. (T-N)
	Excl.	Accom.	Excl. + Accom.	Excl.	Excl.
Nation	2.5	9.7	12.1	7.2	4.7
9-MA	4.0	15.0	19.0	7.9	3.9
8-MN	2.1	8.7	10.8	4.3	2.2
5-CO	0.8	10.0	10.8	4.1	3.3
4-CT	1.3	12.3	13.6	8.5	7.2
7-NC	1.8	12.4	14.2	11.4	9.6
6-IN	2.6	12.2	14.7	6.3	3.7
3-FL	1.8	16.1	18.0	6.9	5.1
2-CA	1.1	7.5	8.5	5.6	4.5
1-AL	1.2	3.6	4.8	4.6	3.4

Note: Excl.=Excluded; Accom.=Accommodated; T-N=TIMSS minus NAEP

Table 60. NAEP and TIMSS Exclusion and Accommodation Rates—Science

2011 NAEP/TIMSS Science: Exclusion & Accommodation Percentages					
State	NAEP			TIMSS	Diff (T-N)
	Excl.	Accom.	Excl. + Accom.	Excl.	Excl.
Nation	1.6	10.6	12.2	7.2	5.6
9-MA	3.2	16.0	19.2	7.9	4.7
8-MN	2.0	8.5	10.4	4.3	2.3
7-CO	0.9	10.3	11.3	4.1	3.2
5-CT	1.3	12.6	13.9	8.5	7.2
6-IN	1.3	12.9	14.2	6.3	5.0
3-FL	1.2	16.3	17.5	6.9	5.7
4-NC	1.6	12.1	13.7	11.4	9.8
2-CA	1.8	7.8	9.5	5.6	3.8
1-AL	1.1	4.1	5.2	4.6	3.5

Note: Excl.=Excluded; Accom.=Accommodated; T-N=TIMSS minus NAEP

Table 61 shows the correlation of errors in estimating TIMSS state scale score means with NAEP and TIMSS exclusion and NAEP accommodation rates. As shown, state differences in the percentage of students accommodated were highly correlated (about .8) with errors in the state mean estimates. Differences in NAEP accommodation rates are significant for two reasons. First, the additional students excluded from the TIMSS assessment are most likely students requiring accommodations in the NAEP assessment that are not provided in TIMSS. Roughly 10 percent of students taking NAEP received accommodations. The percentage of students included in NAEP but not TIMSS was about half of this number. This means that at least half of the students receiving accommodations in NAEP did participate in TIMSS, most likely without these accommodations. Differences in the use of accommodations may also have led to mean score differences for these students. NAEP collects questionnaire data for SD and ELL that provide information about specific student disabilities and characteristics. TIMSS does not collect comparable background information on the students.

Table 61. Correlation of Estimation Error with Exclusion Rate Differences and NAEP Accommodation Rates

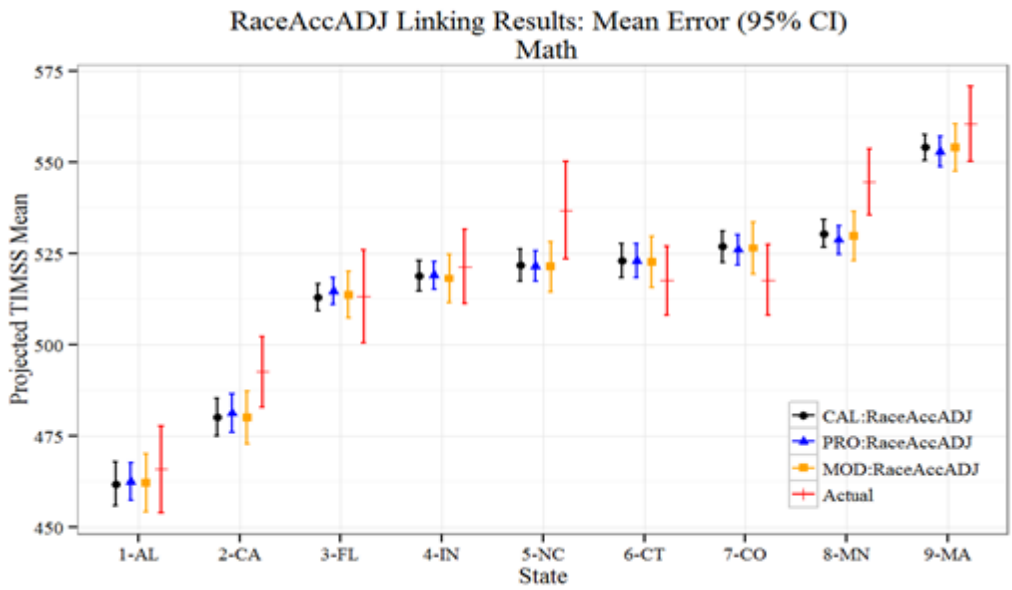
Subject	Method	Correlation with Exclusion Rate Differences (N-T)	Correlation with NAEP Accommodation Rates
Math	MOD	0.39	-0.72
Math	PRO	0.37	-0.74
Math	CAL	0.39	-0.72
Science	MOD	0.45	-0.79
Science	PRO	0.38	-0.81
Science	CAL	0.48	-0.78

Note: Estimation errors were computed as the Predicted TIMSS mean minus the Observed TIMSS mean.
Excl.=Excluded; Accom.=Accommodated; N-T=NAEP minus TIMSS

HumRRO investigated several methods for adjusting the NAEP samples to reduce the impact of differences in exclusion and accommodation rates. The first approach (Accommodations Reweighted, or AccRW) involved proportionally reducing the weight assigned to each student receiving accommodations by an amount related to the difference between the NAEP and TIMSS exclusion rates for each state. The ratio of sum of weights for the reweighted and original NAEP sample was equal to the ratio of the TIMSS and NAEP *inclusion* rates. We also examined options for reweighting accommodated students differentially based on type of accommodation or primary disability code, but found that these options provided essentially the same results as the proportional reweighting. (See full technical report, forthcoming, for more details.)

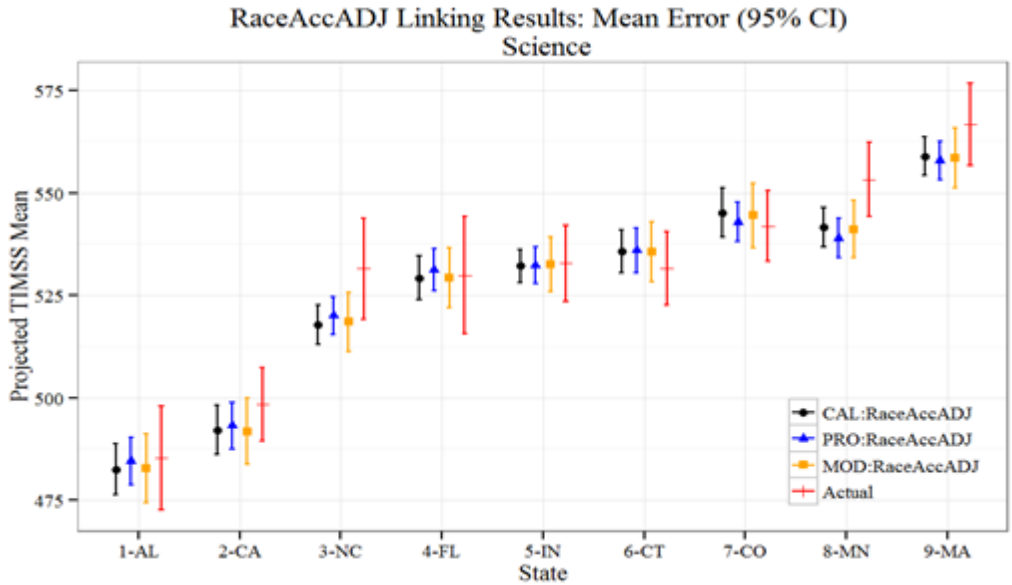
A second approach (Accommodations Adjusted, or AccADJ) involved an empirically derived adjustment based on the percentage of students accommodated in NAEP. We do not have actual TIMSS exclusion rates for states not participating in TIMSS. The percent accommodated in NAEP is the best available predictor of the difference in NAEP and TIMSS exclusion rates. The correlations shown in Table 61 led to an adjustment that added approximately two TIMSS scale score points for every percentage point that a state's NAEP accommodation rate exceeded the national average. Thus, the initial AccADJ model used a coefficient of 2 to compute adjusted predicted means.

In reviewing initial results with our technical panel, it was noted that the race/ethnicity distributions differed for the NAEP and TIMSS samples in several of the validation states. This difference may have resulted from differences in exclusion rate by race/ethnicity or may have resulted from differences in school and class participation rates by race/ethnicity that were not fully accounted for in nonresponse adjustments. A third approach (Race Adjusted, or RaceADJ) involved reweighting the NAEP samples for each state to yield the race/ethnicity distribution of the TIMSS state sample. We also examined an adjustment (Accommodations and Race Adjusted, or RaceAccADJ) that combined the race/ethnicity adjustment and the adjustment based on accommodation rates. Figures 9 and 10 display the adjusted means using the RaceAccADJ. While the RaceAccADJ did improve prediction, it was not feasible to use this approach for states not participating in TIMSS, since TIMSS race/ethnicity distributions would not be available.



Note: The adjustment coefficients and observed percent accommodated were treated as constants so that the standard errors and confidence bounds for the adjusted estimates were the same as for the original estimated TIMSS means.

Figure 9. Adjusted projected TIMSS means using the race and accommodation adjustment (RaceAccADJ) – Mathematics



Note: The adjustment coefficients and observed percent accommodated were treated as constants so that the standard errors and confidence bounds for the adjusted estimates were the same as for the original estimated TIMSS means.

Figure 10. Adjusted projected TIMSS means using the race and accommodation adjustment (RaceAccADJ) – Science

Figure 11 shows the RMSE for estimates of state means using each of the three linkage methods and each of the four adjustments. The race/ethnicity adjustment, by itself, yielded only a small reduction in error. The accommodation adjustment and the combination of race/ethnicity and accommodation adjustments led to the largest reduction in errors.

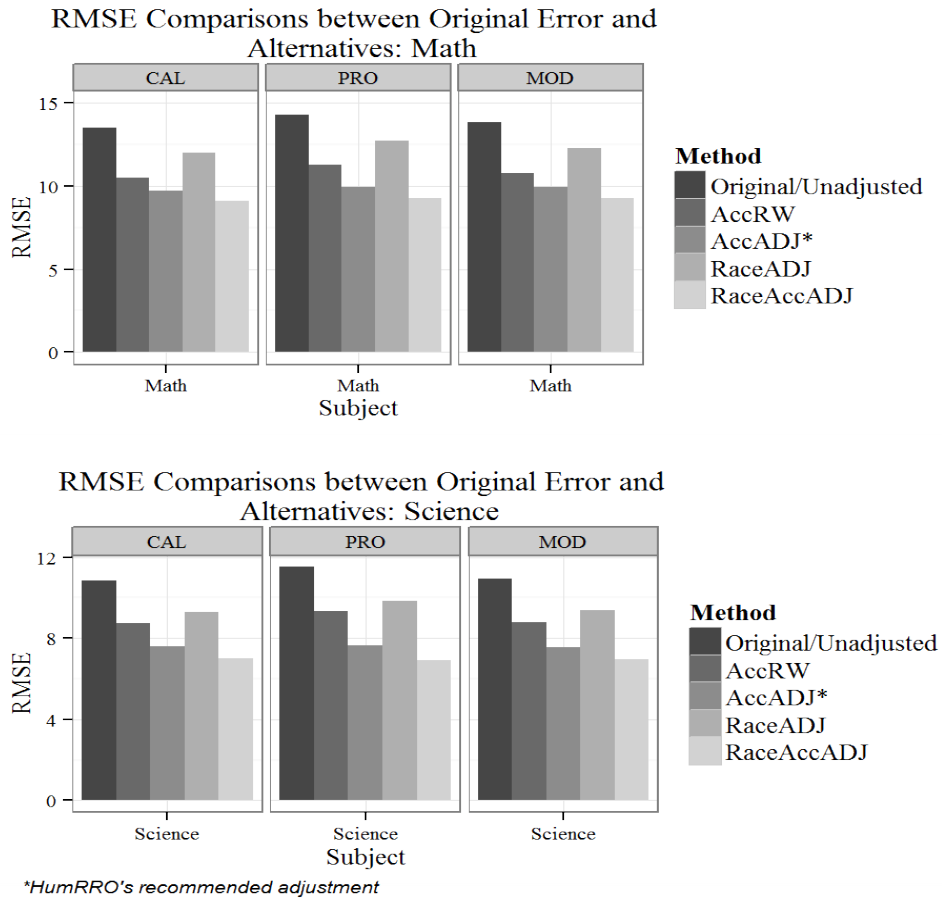


Figure 11. Comparison of error rates resulting from each of the four adjustments for exclusion and accommodation differences.

Finding 4. An adjustment based on the percentage of students accommodated in the NAEP assessment led to a significant reduction in errors in estimating TIMSS scale score means.

Test item differences

Examination of NAEP and TIMSS differences in item difficulty and format did not lead to any plausible corrections to the NAEP-TIMSS linkages. There were no significant differences in the distribution of item difficulties for multiple choice and constructed response items. There were differences in the degree to which the assessments used short and extended constructed response items, but there were no significant differences among states in students' relative performance on

the different item types. A more complete description of these analyses will be provided in our detailed technical report.

Final Accommodation Adjustment

Based on the various accommodation adjustments we examined, we found the empirically derived accommodation adjustment (AccADJ) would result in the state mean predictions with the lowest RMSE[8].

Adjustments to Predicted State Mean Estimates

After adjustment for accommodation and exclusion differences (using AccADJ), differences between the empirical and linkage-based estimates were still larger than could be accounted for by the current estimates of standard errors for the different estimates. Use of this adjustment led to smaller residual errors in comparison to the original, unadjusted linkage-based estimates. For these analyses, we recomputed more precise estimates of the coefficients for the percentage of students receiving NAEP accommodations. Tables 62 and 63 show mean estimates using the revised coefficients of 2.65 for mathematics and 2.21 for science.[9] These coefficients were obtained by regressing the difference between the empirical estimate, $\bar{Y}_j$, and linkage-based estimates, $\bar{Z}_{1j}$, on the difference between the national accommodation rate, A_{NT}, and accommodation rate, A_j, for each of the nine validation states j, using a model without an intercept:

$$\left(\bar{Y}_j - \bar{Z}_{1j}\right) = \beta_{adj}\left(A_j - A_{NT}\right) + e_j \qquad \text{(H1)}$$

where the adjustment coefficient is estimated by

$$\hat{\beta}_{adj} = \frac{\sum_j \left(A_j - A_{NT}\right)\left(\bar{Y}_j - \bar{Z}_{1j}\right)}{\sum_j \left(A_j - A_{NT}\right)^2} \qquad \text{(H2)}$$

These coefficients, as well as the national NAEP accommodation rates documented in Tables 59 and 60 (9.7 percent in mathematics and 10.6 percent in science), were provided to AIR to inform the adjustments described on page 33.

[8] Although the RMSE was lowest for the RaceAccADJ, because we would not be able to apply that to all 50 states, we did not view this as a viable adjustment.

[9] Projections in Tables 22 and 24 were calculated from accommodation adjusted mean estimates using step 3 of the moderation approach described earlier and differ from projections in Tables 62 and 63.

Table 62. Predicted State Mean Estimates For the Statistical Moderation Using AccADJ—Mathematics

State	Actual TIMSS	Mean Estimates using:		Projected Error (P - A)	
		Unadj.	AccADJ	Unadj.	AccADJ
Nation	506.89	506.89	-	0.00	-
9-MA	560.58	540.00	554.07	-20.58	-6.51
8-MN	544.73	532.52	529.80	-12.21	-14.93
7-CO	517.79	525.80	526.52	8.00	8.73
6-CT	517.62	515.85	522.68	-1.78	5.06
5-NC	536.90	514.31	521.41	-22.59	-15.49
4-IN	521.51	511.66	518.22	-9.85	-3.29
3-FL	513.30	496.63	513.72	-16.68	0.42
2-CA	492.62	486.00	480.13	-6.62	-12.49
1-AL	465.93	478.30	462.14	12.37	-3.79
			Root Mean Square Error:	**13.83**	**9.93**

Note: Unadj.=Unadjusted; AccADJ=Accommodation Adjustment

Table 63. Predicted State Mean Estimates For the Statistical Moderation Using AccADJ—Science

State	Actual TIMSS	Mean Estimates using:		Projected Error (P - A)	
		Unadj.	AccADJ	Unadj.	AccADJ
Nation	522.19	522.19	-	0.00	-
9-MA	566.78	546.63	558.53	-20.15	-8.25
8-MN	553.27	545.86	541.18	-7.41	-12.09
7-CO	541.95	545.12	544.50	3.17	2.55
6-CT	531.60	531.34	535.68	-0.26	4.08
5-IN	532.80	527.35	532.51	-5.45	-0.30
4-FL	529.89	516.71	529.29	-13.18	-0.60
3-NC	531.53	515.16	518.54	-16.37	-12.99
2-CA	498.52	498.12	491.86	-0.40	-6.66
1-AL	485.37	497.10	482.80	11.73	-2.57
			Root Mean Square Error:	**10.95**	**7.56**

Note: Unadj.=Unadjusted; AccADJ=Accommodation Adjustment

Adjustments to Standard Error Estimates

Additional analyses performed as part of this evaluation involved developing an estimate of the additional variance in linkage-based estimates. The approach used will be described in detail in the forthcoming technical report for this linking study. Briefly, it involved examining the variance of differences between the empirical and linkage-based estimates of state means and subtracting out known estimates of variance due to NAEP and TIMSS sampling and measurement error and linkage error.

Tables 64 and 65 show estimates of the different NAEP and TIMSS variance components for each state and the squared difference between the linkage-based and empirical TIMSS mean estimates for the state. These analyses used the linkage derived from statistical moderation, because the assumptions of this model are fewer and the linkage error variance is well estimated for this method. Also, we used the predicted TIMSS mean estimates that included the accommodation adjustment described above.

We averaged the variance component estimates across the nine validation states and then subtracted these variance components from an unbiased estimate of the mean squared error using eight degrees of freedom to get an unbiased estimate of residual error. This residual error is a consequence of the various differences between the two assessments, although we cannot attribute specific amounts of variance to specific differences. We have labeled this residual variation as “model error” to indicate that the variance results from differences in the two assessment models. Tables 66 and 67 show the impact of adding model error into standard error estimates for the linkage-based state means. Further analyses indicated that none of the differences between linkage-based and empirical estimates of TIMSS state means were statistically significant when the expanded standard error estimates were used. Figures 12 and 13 display the accommodation adjustment means (AccADJ) for math and science with model error.

Table 64. Estimation of Model Error Variation for the AccADJ Statistical Moderation Linkage – Mathematics

State	Total Error		Variances in MOD Estimates				Variances in TIMSS		
	Error	Error^2	Total	Sample	Meas.	Link.	Total	Samp.	Meas.
MA	-6.51	42.36	10.82	2.78	0.11	7.93	27.86	27.52	0.34
MN	-14.93	222.76	11.92	3.63	0.58	7.71	21.25	21.1	0.16
CO	8.73	76.2	12.89	5.16	0.17	7.56	24.01	23.04	0.97
CT	5.06	25.61	12.63	4.71	0.5	7.43	23.45	22.78	0.67
NC	-15.49	239.92	11.92	4.33	0.18	7.41	46.91	45.95	0.96
IN	-3.29	10.83	11.72	4.02	0.3	7.4	26.32	25.73	0.59
FL	0.42	0.17	10.56	2.98	0.15	7.44	41.57	41.46	0.11
CA	-12.49	156.01	13.95	6.21	0.14	7.6	23.84	23.44	0.4
AL	-3.79	14.35	16.41	7.83	0.79	7.79	36.68	36.05	0.63
MSE=		**98.53**	**12.54**				**30.21**		
Model Error =		**55.78**	**(Total error - NAEP Estimate Variance - TIMSS Variance)**						

Note: MOD=Moderation; Meas.=Measurement; Link.=Linking; MSE=Mean Square Error

Table 65. Estimation of Model Error Variation for the AccADJ Statistical Moderation Linkage – Science

State	Total Error		Variances in MOD Estimates				Variances in TIMSS		
	Error	Error^2	Total	Sample	Meas.	Link.	Total	Sample	Meas.
MA	-8.25	68.06	13.92	4.81	1.82	7.3	26.24	25.71	0.54
MN	-12.09	146.16	12.91	5.15	0.48	7.28	21.57	21.05	0.53
CO	2.55	6.53	16.09	8.6	0.21	7.27	19.39	18.2	1.19
CT	4.08	16.66	13.89	6.17	0.62	7.1	20.85	20.75	0.1
IN	-0.3	0.09	11.51	3.65	0.78	7.08	22.6	21.63	0.97
FL	-0.6	0.35	14.04	6.66	0.3	7.08	53.36	50.96	2.4
NC	-12.99	168.87	13.39	5.91	0.4	7.09	39.42	38.69	0.73
CA	-6.66	44.34	16.81	8.62	0.9	7.29	20.75	19.97	0.79
AL	-2.57	6.6	18.05	9.93	0.8	7.31	41.72	39.78	1.94
	MSE=	**57.21**	**14.51**				**29.55**		
Model Error =		**13.15**	**(Total error - NAEP Estimate Variance - TIMSS Variance)**						

Note: MOD=Moderation; Meas.=Measurement; Link.=Linking; MSE=Mean Square Error

Table 66. Statistical Significance of Differences in Estimates of TIMSS Scale Score Means for Unadjusted (Without Model Error) and Adjusted Means (With Model Error) for the Statistical Moderation Linkage – Mathematics

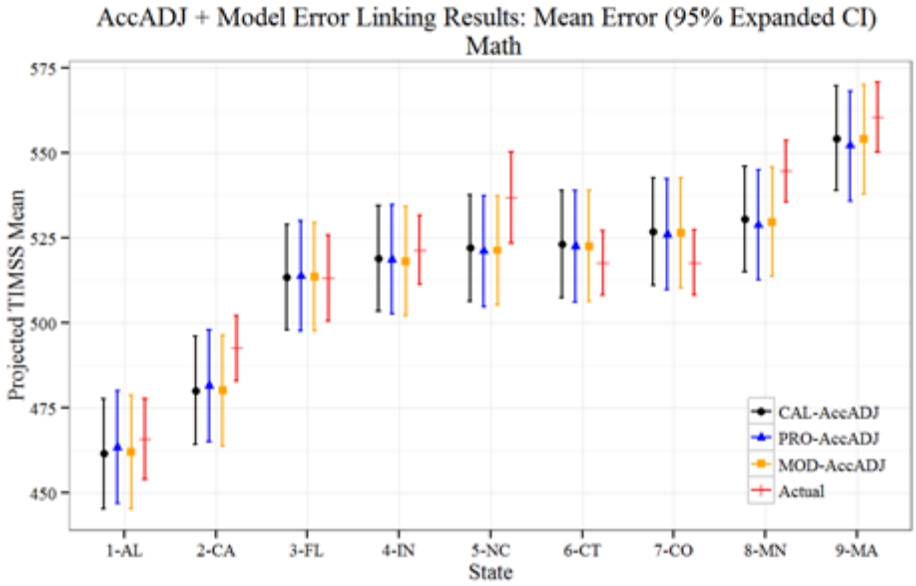
			Unadjusted With No Model Error				AccADJ With Model Error			
	Actual TIMSS		Projected		Error		Projected		Error	
State	Mean	SE	Mean	SE	Diff.	t	Mean	SE	Diff.	t
Nation	506.89	2.63	506.89	2.75	0.00	0.00	-	-	-	-
9-MA	560.58	5.28	**540.00**	**3.29**	**20.58**	**-3.31**	554.07	8.16	6.51	-0.67
8-MN	544.73	4.61	**532.52**	**3.45**	**12.21**	**-2.12**	529.80	8.23	14.93	-1.58
7-CO	517.79	4.9	525.80	3.59	-8.00	1.32	526.52	8.29	-8.73	0.91
6-CT	517.62	4.84	515.85	3.55	1.78	-0.30	522.68	8.27	-5.06	0.53
5-NC	536.9	6.85	**514.31**	**3.45**	**22.59**	**-2.94**	521.41	8.23	15.49	-1.45
4-IN	521.51	5.13	511.66	3.42	9.85	-1.60	518.22	8.22	3.29	-0.34
3-FL	513.3	6.45	**496.63**	**3.25**	**16.68**	**-2.31**	513.72	8.15	-0.42	0.04
2-CA	492.62	4.88	486.00	3.73	6.62	-1.08	480.13	8.35	12.49	-1.29
1-AL	465.93	6.06	478.30	4.05	-12.37	1.70	462.14	8.50	3.79	-0.36

Note: Bold font indicates predicted means are statistically significant from the actual means.

Table 67. Statistical Significance of Differences in Estimates of TIMSS Scale Score Means for Unadjusted (Without Model Error) and Adjusted Means (With Model Error) for the Statistical Moderation Linkage – Science

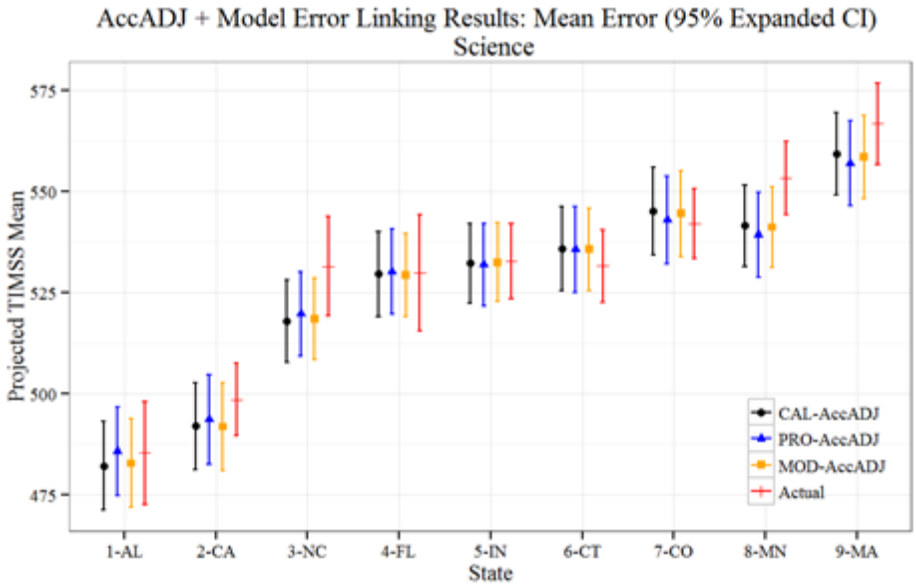
			Unadjusted With No Model Error				AccADJ With Model Error			
	Actual TIMSS		Projected		Error		Projected		Error	
State	Mean	SE	Mean	SE	Diff.	t	Mean	SE	Diff.	t
Nation	522.19	2.53	522.19	2.71		0.00	-	-	-	-
9-MA	566.78	5.12	**546.63**	**3.73**	**20.15**	**-3.18**	558.53	5.20	8.25	-1.13
8-MN	553.27	4.64	545.86	3.59	7.41	-1.26	541.18	5.10	12.09	-1.75
7-CO	541.95	4.4	545.12	4.01	-3.17	0.53	544.50	5.41	-2.55	0.37
6-CT	531.6	4.57	531.34	3.73	0.26	-0.04	535.68	5.20	-4.08	0.59
5-IN	532.8	4.75	527.35	3.39	5.45	-0.93	532.51	4.97	0.30	-0.04
4-FL	529.89	7.3	516.71	3.75	13.18	-1.61	529.29	5.21	0.60	-0.07
3-NC	531.53	6.28	515.16	3.66	16.37	-2.25	518.54	5.15	12.99	-1.60
2-CA	498.52	4.56	498.12	4.10	0.40	-0.07	491.86	5.47	6.66	-0.94
1-AL	485.37	6.23	497.10	4.25	-11.73	1.52	482.80	5.59	2.57	-0.30

Bold font indicates predicted estimates are statistically significant from the actual estimates.



*Note: We did not create a separate AccADJ equation for the CAL and PRO methods, but we did compute "residual error" separately for each method and created confidence bounds that reflected the revised total error estimates.

Figure 12. Adjusted projected TIMSS means using the accommodation adjustment (AccADJ) and incorporating model error in the confidence bands – Mathematics



*Note: We did not create a separate AccADJ equation for the CAL and PRO methods, but we did compute "residual error" separately for each method and created confidence bounds that reflected the revised total error estimates.

Figure 13. Adjusted projected TIMSS means using the accommodation adjustment (AccADJ) and incorporating model error in the confidence bands – Science

Adjustments to Percent Above Cut Estimates

HumRRO investigated two approaches for adjusting percent above (Benchmark Level) cut estimates using empirical adjustment based on the percentage of students accommodated in NAEP. The first approach is based on a normal approximation to the projected TIMSS score distribution (AccADJ_Normal). In this approach, we first converted the original percent estimate into TIMSS scale score metric using the inverse normal cumulative distribution with mean equal to the unadjusted TIMSS mean estimate. This gives a "normalized" cut score that may differ from the original cut score depending on how the predicted TIMSS score distribution differs from a normal distribution. The adjusted percent above cut estimate was then obtained by evaluating the cumulative normal distribution with mean equal to the TIMSS mean estimate that included the empirically derived accommodation adjustment at the normalized cut score. To estimate the standard error for the adjusted percent above cut estimate, we added and subtracted the adjusted estimate of the standard error of the TIMSS mean estimate that included model error (see Table 64). The standard error estimate is half of the difference between their corresponding percentiles based on the normal distribution with adjusted mean.

The second approach applied the accommodation adjustment method directly using the percentile metric by regressing the percent above cut prediction error on NAEP accommodation rates (AccADJ_Direct). In this approach the adjustment coefficient for the percentage of students receiving NAEP accommodations was estimated separately by benchmark level. The adjusted predicted percent above cut estimates are obtained by adding the adjustment to the original predicted percent above cut estimate. Corresponding adjusted standard errors were obtained in the same fashion as in accommodation adjustment for the mean. We obtained an unbiased estimate of the mean squared error by dividing the sum of the squared difference between adjusted NAEP predicted and empirical TIMSS percent above cut estimates by eight degrees of freedom. We then averaged the NAEP and TIMSS variance components for percent above cut scores across the nine validation states and subtracted these from unbiased estimates of the mean squared error to get an estimate model error. The adjusted standard error for NAEP predicted percent above cut estimate is the square root of the sum of the model error and the original variance.

Table 68 shows the mean squared errors (MSEs) for the unadjusted NAEP projected percent above cut estimates and the two adjusted estimates. Tables 69 through 72 compare the two percent above cut adjustments (AccADJ_Normal and AccADJ_Direct) with model error with the unadjusted estimates without model error. As seen in Tables 69 through 72, the AccADJ_Direct results in one negative estimate (Table 72) and negative model errors for three of the benchmark levels (Tables 69-71). These results, combined with comparisons of the MSEs suggest that normal approximation adjustment is as good as or better than the direct adjustment across all benchmark levels. The normal approximation is also the more parsimonious method because it only requires one adjustment equation as opposed to four separate adjustment equations by benchmark level used in the direct approach. For these reasons, we recommend the normal approximation method to adjust the percent above cut estimates for differences in NAEP accommodation rates.

Table 68. MSEs for Unadjusted (Without Model Error) and Adjusted (With Model Error) Percent Above Cut Estimates for the Statistical Moderation Linkage

	Math			Science		
Cut Score	Unadj.	AccADJ _Normal	AccADJ _Direct	Unadj.	AccADJ_ Normal	AccADJ _Direct
>=400	9.46	5.24	7.03	4.28	2.87	3.19
>=475	27.17	16.74	15.05	14.49	7.24	7.65
>=550	47.21	24.32	29.54	19.94	9.62	10.07
>=625	17.49	9.14	10.73	22.77	15.15	16.95

* Unadj. - No adjustment for % ACC
AccADJ_Normal - Using adjustment to the mean and the normal approximation
AccADJ_Direct - Direct adjustment using a separate regression equation for each cutoff

Table 69. Statistical Significance of Differences in Estimates of Percent Above Low TIMSS Benchmark Level Cutoffs for the Unadjusted (Without Model Error) and Adjusted (With Model Error) Statistical Moderation Approaches

Mathematics			Unadj. (Without Model Error)			AccADJ_Normal (With Model Error)			AccADJ_Direct (With Model Error)		
	Actual TIMSS		Projected		Error	Projected		Error	Projected		Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	97.72	0.34	96.62	0.67	-1.1	97.85	0.59	0.13	99.10	1.96	1.38
8-MN	97.17	0.67	95.32	0.86	-1.85	94.95	1.17	-2.23	94.84	2.04	-2.33
7-CO	93.48	1.07	94.67	1.03	1.18	94.77	1.18	1.29	94.79	2.11	1.31
6-CT	90.72	1.43	93.92	1.12	3.2	**94.95**	**1.16**	**4.24**	95.12	2.16	4.41
5-NC	95.34	1.31	93.42	1.24	-1.91	94.55	1.21	-0.79	94.67	2.22	-0.67
4-IN	95.07	0.96	94.49	1.18	-0.59	95.47	1.14	0.40	95.64	2.19	0.56
3-FL	93.76	1.31	90.32	1.41	-3.44	93.73	1.37	-0.03	93.32	2.33	-0.44
2-CA	87.45	1.72	85.36	1.84	-2.1	83.67	2.49	-3.79	84.32	2.61	-3.13
1-AL	78.61	2.32	**85.59**	**2.08**	**6.97**	80.11	3.17	1.50	82.74	2.78	4.13

Science			Unadj. (Without Model Error)			AccADJ_Normal (With Model Error)			AccADJ_Direct (With Model Error)		
	Actual TIMSS		Projected		Error	Projected		Error	Projected		Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	96.47	0.66	95.45	0.85	-1.02	96.71	0.48	0.24	97.17	**0.85**	0.70
8-MN	97.83	0.7	96.16	0.77	-1.67	**95.60**	**0.65**	**-2.23**	**95.48**	**0.77**	**-2.35**
7-CO	96.31	0.68	95.69	0.85	-0.62	95.62	0.66	-0.69	95.60	**0.85**	-0.71
6-CT	92.05	1.28	93.69	1.02	1.65	94.34	0.75	2.30	94.32	**1.02**	2.27
5-IN	95.11	0.86	94.28	1	-0.82	95.04	0.68	-0.07	95.03	**1.00**	-0.08
4-FL	93.48	1.49	91.28	1.39	-2.2	93.50	0.82	0.03	93.10	**1.39**	-0.38
3-NC	94.37	1.38	92.13	1.39	-2.25	92.74	0.91	-1.63	92.61	**1.39**	-1.76
2-CA	87.53	1.64	86.13	1.78	-1.4	84.48	1.49	-3.04	85.22	**1.78**	-2.30
1-AL	83.39	1.91	87.76	2.04	4.36	83.87	1.66	0.47	85.69	**2.04**	2.30

Note: Bold font indicates predicted estimates are statistically significant from the actual estimates.
Bold underlined font indicates that the model error was negative; thus, the SE estimates were set to equal the unadjusted SEs.

Table 70. Statistical Significance of Differences in Estimates of Percent Above Intermediate TIMSS Benchmark Level Cutoffs for the Unadjusted (Without Model Error) and Adjusted (With Model Error) Statistical Moderation Approaches

Mathematics			Unadj. (Without Model Error)			AccADJ_Normal (With Model Error)			AccADJ_Direct (With Model Error)		
	Actual TIMSS		Projected		Error	Projected		Error	Projected		Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	88.07	1.39	**82.04**	**1.96**	**-6.04**	86.69	2.43	-1.38	87.17	2.43	-0.90
8-MN	82.75	1.86	79.44	2.17	-3.31	78.38	3.27	-4.38	78.45	2.61	-4.30
7-CO	70.58	2.53	75.87	2.43	5.29	76.17	3.41	5.59	76.13	2.83	5.55
6-CT	69.25	2.55	70.4	2.62	1.15	73.50	3.65	4.25	72.89	2.99	3.64
5-NC	77.90	2.51	**70.17**	**2.44**	**-7.73**	73.36	3.59	-4.54	72.76	2.83	-5.14
4-IN	74.13	2.34	71.16	2.67	-2.97	74.32	3.84	0.20	73.55	3.04	-0.58
3-FL	67.60	3.31	62.2	2.5	-5.4	70.66	3.81	3.06	68.43	2.89	0.83
2-CA	59.04	2.76	56.11	2.68	-2.93	53.30	4.00	-5.74	53.97	3.05	-5.07
1-AL	45.76	3.2	53.6	3.15	7.85	44.99	4.49	-0.77	47.71	3.47	1.96

Science			Unadj. (Without Model Error)			AccADJ_Normal (With Model Error)			AccADJ_Direct (With Model Error)		
	Actual TIMSS		Projected		Error	Projected		Error	Projected		Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	87.09	1.54	82.82	2.03	-4.27	86.36	1.43	-0.73	86.79	**2.03**	-0.30
8-MN	85.39	2.02	83.94	2.03	-1.46	82.34	1.79	-3.05	82.38	**2.03**	-3.02
7-CO	79.59	1.96	82.58	2.22	2.99	82.37	1.84	2.78	82.38	**2.22**	2.79
6-CT	74.23	2	77.66	2.4	3.43	79.27	1.89	5.04	79.10	**2.40**	4.87
5-IN	77.72	2.09	76.83	2.31	-0.89	78.89	1.93	1.17	78.55	**2.31**	0.83
4-FL	73.83	3.55	71.14	2.52	-2.7	76.23	2.00	2.39	75.33	**2.52**	1.50
3-NC	74.90	2.98	71.88	2.54	-3.02	73.32	2.16	-1.58	73.01	**2.54**	-1.89
2-CA	62.03	2.54	63.26	2.7	1.23	60.54	2.40	-1.49	61.17	**2.70**	-0.86
1-AL	56.20	3.73	64.61	3.08	8.41	57.96	2.66	1.76	59.84	**3.08**	3.64

Note: Bold indicates predicted estimates are statistically significant from the actual estimates.
Bold underlined font indicates that the model error was negative; thus, the SE estimates were set to equal the unadjusted SEs.

Table 71. Statistical Significance of Differences in Estimates of Percent Above High TIMSS Benchmark Level Cutoffs for the Unadjusted (Without Model Error) and Adjusted (With Model Error) Statistical Moderation Approaches

Mathematics			Unadj. (Without Model Error)			AccADJ_Normal (With Model Error)			AccADJ_Direct (With Model Error)		
	Actual TIMSS		Projected		Error	Projected		Error	Projected		Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	57.35	3.22	**46.28**	**2.64**	**-11.07**	54.06	4.48	-3.29	52.61	4.61	-4.74
8-MN	48.90	2.84	42.55	2.73	-6.35	41.11	4.33	-7.79	41.33	4.66	-7.57
7-CO	35.14	2.69	38.7	2.69	3.56	39.07	4.22	3.93	39.03	4.64	3.89
6-CT	36.52	2.94	33.33	2.67	-3.2	36.74	4.19	0.22	36.40	4.63	-0.12
5-NC	44.24	3.6	**32.4**	**2.48**	**-11.84**	35.86	4.08	-8.37	35.59	4.52	-8.65
4-IN	35.32	3.33	29.51	2.64	-5.81	32.88	4.31	-2.44	32.45	4.61	-2.87
3-FL	31.11	3.16	23.69	2.16	-7.41	31.44	3.93	0.33	31.37	4.36	0.27
2-CA	24.40	2.46	21.72	2.14	-2.68	19.69	2.80	-4.71	19.08	4.35	-5.32
1-AL	14.73	2.55	16.51	2.28	1.78	11.70	2.24	-3.02	9.25	4.42	-5.48

Science			Unadj. (Without Model Error)			AccADJ_Normal (With Model Error)			AccADJ_Direct (With Model Error)		
	Actual TIMSS		Projected		Error	Projected		Error	Projected		Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	61.46	2.79	**52.9**	**2.86**	**-8.57**	58.79	2.54	-2.68	57.64	**2.86**	-3.82
8-MN	53.67	2.62	52.23	3.04	-1.44	49.71	2.76	-3.96	50.37	**3.04**	-3.30
7-CO	47.86	2.58	51.34	3.35	3.48	51.02	2.84	3.16	51.10	**3.35**	3.24
6-CT	44.97	2.47	44.18	2.75	-0.79	46.36	2.62	1.39	45.91	**2.75**	0.94
5-IN	43.37	2.85	41.82	2.61	-1.55	44.54	2.63	1.17	43.88	**2.61**	0.51
4-FL	41.52	3.46	36.86	2.67	-4.65	42.89	2.54	1.37	41.88	**2.67**	0.36
3-NC	42.22	3.2	34.84	2.58	-7.37	36.45	2.47	-5.77	36.19	**2.58**	-6.03
2-CA	28.09	1.94	29.31	2.41	1.23	26.91	2.06	-1.18	26.82	**2.41**	-1.27
1-AL	23.77	2.76	27.14	2.51	3.37	21.69	2.00	-2.07	21.44	**2.51**	-2.33

Note: Bold font indicates predicted estimates are statistically significant from the actual estimates.
Bold underlined font indicates that the model error was negative; thus, the SE estimates were set to equal the unadjusted SEs.

Table 72. Statistical Significance of Differences in Estimates of Percent Above Advanced TIMSS Benchmark Level Cutoffs for the Unadjusted (Without Model Error) and Adjusted (With Model Error) Statistical Moderation Approaches

Mathematics			Unadj. (Without Model Error)			AccADJ_Normal (With Model Error)			AccADJ_Direct (With Model Error)		
	Actual TIMSS		Projected		Error	Projected		Error	Projected		Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	19.26	2.97	**11.33**	**1.69**	**-7.93**	15.53	2.70	-3.73	15.23	2.94	-4.04
8-MN	13.08	2.31	9.84	1.6	-3.25	9.21	1.85	-3.87	9.08	2.89	-4.00
7-CO	7.7	1.14	8.73	1.55	1.03	8.88	1.77	1.18	8.93	2.87	1.23
6-CT	10.17	1.34	6.93	1.31	-3.24	8.25	1.70	-1.92	8.83	2.74	-1.34
5-NC	13.75	2.63	**6.93**	**1.32**	**-6.82**	8.28	1.67	-5.47	8.89	2.75	-4.86
4-IN	6.98	1.18	4.38	1.12	-2.61	5.34	1.30	-1.65	6.19	2.66	-0.79
3-FL	7.92	1.59	**3.58**	**0.83**	**-4.34**	5.83	1.30	-2.08	8.31	2.55	0.39
2-CA	4.82	0.91	4.4	1.06	-0.41	3.78	0.83	-1.04	2.78	2.63	-2.04
1-AL	2.1	0.77	1.91	0.67	-0.19	1.10	0.33	-1.00	-2.56	2.50	-4.66

Science			Unadj. (Without Model Error)			AccADJ_Normal (With Model Error)			AccADJ_Direct (With Model Error)		
	Actual TIMSS		Projected		Error	Projected		Error	Projected		Error
State	Est	SE	Est	SE	(P-A)	Est	SE	(P-A)	Est	SE	(P-A)
9-MA	24.46	2.55	**14.74**	**2.06**	**-9.72**	18.45	1.74	-6.00	18.71	3.93	-5.75
8-MN	16.13	1.87	12.49	1.97	-3.64	**11.23**	**1.32**	**-4.90**	10.93	3.89	-5.20
7-CO	14.46	1.62	13.68	2.09	-0.78	13.50	1.55	-0.95	13.48	3.95	-0.98
6-CT	14.07	1.54	10.31	1.76	-3.76	11.34	1.27	-2.73	11.76	3.79	-2.31
5-IN	10.42	1.35	7.22	1.48	-3.2	8.22	1.01	-2.20	8.94	3.67	-1.48
4-FL	13.32	1.97	**7.34**	**1.34**	**-5.98**	9.77	1.12	-3.55	11.53	3.61	-1.78
3-NC	12.42	2.18	**6.51**	**1.34**	**-5.92**	**7.07**	**0.89**	**-5.35**	7.63	3.61	-4.79
2-CA	6.03	0.73	5.76	1.32	-0.27	4.98	0.64	-1.05	3.67	3.60	-2.36
1-AL	4.81	1.01	3.64	1.14	-1.17	**2.45**	**0.39**	**-2.36**	-1.13	3.54	-5.94

Note: Bold font indicates predicted estimates are statistically significant from the actual estimates.
Bold underlined font indicates that the model error was negative; thus, the SE estimates were set to equal the unadjusted SEs.

Recommendations

Recommendation 1: Use estimates from the statistical moderation linkages.

SUPPORTING REFERENCES:
- Tables 41 – 58
- Figures 5 – 6

While results indicated slight improvements in estimates using the joint calibration (CAL) approach, the differences do not justify the extra effort and expense associated with this approach in future years. In addition, we have not fully investigated all of the assumptions of the joint calibration approach, most notably the stability of item parameter estimates across test administration conditions.

Recommendation 2: Use the adjustment based on percent of students accommodated to improve linkage-based estimates.

SUPPORTING REFERENCES:
- Tables 59 – 63
- Figure 11

The other adjustments examined were useful in understanding the impact of test administration differences, but cannot be used in situations where TIMSS exclusion rates or race/ethnicity distributions are not available. The adjustment based only on the NAEP accommodation rate did lead to a reduction in differences between the linkage-based and empirical estimates of state means.

Recommendation 3: Include an estimate of model error in standard error estimates and confidence bounds for linkage-based estimates.

SUPPORTING REFERENCES:
- Tables 44 – 57
- Figures 11 and 12

Accurate confidence bounds are critical to supporting valid conclusions about linkage-based estimates. Additional analyses were required to estimate model error when the accommodation adjustment is used. Additional analyses to estimate model error variance for statistics other than state means (such as the percentage of students scoring at or above a TIMSS benchmark level) were also needed. These analyses were subsequently performed by AIR, taking into account the additional projection methodology described above.

Recommendation 4: Use normal approximations to adjust estimates of percent above cut points for consistency with the adjustment based on percent of students accommodated for state mean estimates.

SUPPORTING REFERENCES:

- Tables 68 – 72

As described above, the normal approximation approach avoided negative estimates of model error and was more parsimonious in that it used the same adjustment equation as the TIMSS mean score estimates.

Recommendation 5: Include confidence bounds in all reporting

While some adjustments presented here reduced the confidence intervals from their initial size, the remaining error estimates and confidence intervals are not trivial. The results of this linking could easily be misinterpreted if only point estimates of mean scale scores or percentages of students at or above a benchmark level cutpoint are presented. Readers could construe differences among states or between states and countries where no true differences exist. We strongly encourage the inclusion of confidence intervals and/or error estimates in all reporting to minimize misinterpretation of the information by end users.

References

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 Technical Report* (NCES 2001-509). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Beaton, A. E. (1987). Implementing the new design: The NAEP 1983-84 technical report (NO. 15-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Feuer, M. J., Holland, P. W., Green, G. F., Bertenthal, M. W., & Hemphill, F. C. (1999). Uncommon measures: Equivalence and linkage among educational tests (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). Washington, DC: National Academy Press.

Foy, P., Brossman, B., & Galia, J. (2012). Scaling the TIMSS and PIRLS 2011 Achievement Data. In M. O. Martin, & I. V. Mullis (Eds.), Methods and procedures in TIMSS and PIRLS 2011. Chestnut Hill, MA: Boston College. Retrieved August 19, 2013, from http://timss.bc.edu/methods/pdf/TP11_Scaling_Achievement.pdf

Foy, P., Galia, J., & Li, I. (2008). Scaling the Data from the TIMSS 2007 Mathematics and Science Assessments. In J. F. Olson, M. O. Martin, & I. V. Mullis (Eds.), TIMSS 2007 Technical Report (pp. 225-280). Chestnut Hill, MA: Boston College. Retrieved August 19, 2013, from http://timss.bc.edu/timss2007/PDF/T07_TR_Chapter11.pdf

Haberman, S. J. (2011). Using Exponential Families for Equating. In A. A. von Davier (ed.), Statistical Models for Test Equating, Scaling, and Linking, Statistics for Social and Behavioral Sciences (pp. 125-140). Springer, LLC.

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

Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking. New York, NY: Springer.

Linn, R. L. (1993). Linking results of distinct assessments. Applied Measurement in Education, 6, 83-102.

Lord, F. M., & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. MA: Addison-Wesley.

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

Mislevy, R. J., Beaton, A. E., Kaplan, B. A., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. Journal of Educational Measurement, 29, 122-161.

Mislevy, R., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. Journal of Educational and Behavioral Statistics, 17, 131−154.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Neidorf, T. S., Binkley, M., Gattis, K., & Nohara, D. (2006). Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments (NCES 2006-029). National Center for Education Statistics, U.S. Department of Education. Washington, DC. Retrieved August 19, 2013, from http://nces.ed.gov/pubsearch.

Nohara, D. (2001). A comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Program for International Student Assessment (PISA) (NCES 2001–07). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Provasnik, S., Kastberg, D., Ferraro, D., Lemanski, N., Roey, S., & Jenkins, F. (2012). Highlights From TIMSS 2011: Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context (NCES 2013-009). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? In M. von Davier and D. Hastedt (Eds.), IERI monograph series: Issues and methodologies in large scale assessments (vol. 2). 9-36. IEA-ETS Research Institute. Retrieved August 19, 2013, from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf

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